The use of propensity score matching in the evaluation of active labour market policies.

Alex Bryson
Richard Dorsett
Susan Purdon

Policy Studies Institute and National Centre for Social Research

This is a reproduction of Department for Work and Pensions Research Working Paper Number 4, ISBN 1843880431.

The report is available online:

http://www.dwp.gov.uk/asd/asd5/WP4.pdf

Working Paper Number 4

THE USE OF PROPENSITY SCORE MATCHING IN THE EVALUATION OF ACTIVE LABOUR MARKET POLICIES

THE USE OF PROPENSITY SCORE MATCHING IN THE EVALUATION OF ACTIVE LABOUR MARKET POLICIES

A study carried out on behalf of the Department for Work and Pensions

By

Alex Bryson, Richard Dorsett and Susan Purdon
Policy Studies Institute and National Centre for Social Research

First Published 2002

ISBN 1 84388 043 1
ISSN 1476 3583

Acknowledgements

We would like to thank Liz Rayner, Mike Daly and other analysts in the Department for Work and Pensions for their advice, assistance and support in writing this paper. We are also particularly grateful to Professor Jeffrey Smith at the University of Maryland for extensive and insightful comments on previous versions of this paper.

The authors

Alex Bryson is a Principal Research Fellow at the Policy Studies Institute. He has recently been involved in the evaluations of the New Deal for Young People and ONE and is currently evaluating the impact of tax credits on employers.

Richard Dorsett is also a Principal Research Fellow at the Policy Studies Institute. He has recently been involved in evaluations of the New Deal for Young People, Joint Claims for JSA, New Deal for Partners and is currently evaluating Work-Based Learning for Adults.

Susan Purdon is the Director of the Survey Methods Centre at the National Centre for Social Research. She is currently involved in the evaluation of the National New Deal for Lone Parents and feasibility work for the evaluation of a job retention and rehabilitation pilot.

Acronyms used in this paper

All acronyms are explained in the text.

ATE: average treatment effect
CAB: Citizens' Advice Bureaux
CIA: conditional independence assumption
CSR: common support requirement
DiD: difference-in-differences
EGW: extended Gateway
EMP: employment subsidy under the New Deal for Young People
ETF: Environment Task Force
FTET: full-time education and training option under the New Deal for Young People
IV: instrumental variable
JSA: Jobseeker's Allowance
LATE: local average treatment effect
NDLP: New Deal for Lone Parents
NDLTU: New Deal for the Long Term Unemployed
NDYP: New Deal for Young People
PSM: propensity score matching
TT: treatment on the treated
VS: voluntary sector option under the New Deal for Young People

Contents

1. Introduction

The purpose of labour market policy evaluation is to assess whether changes in policy, and the introduction of new programmes, influence outcomes such as employment and earnings for those subject to the policy change. As one analyst has recently noted: 'The task of evaluation research lies in devising methods to reliably estimate [the impact of policy change], so that informed decisions about programme expansion and termination can be made' (Smith, 2000: 1).

Although UK governments have always intervened in labour markets in pursuit of desirable policy objectives, the evaluation of avowedly active labour market programmes began in earnest with the Restart programme in the late 1980s. Restart required unemployed claimants to attend regular work-focused interviews with a personal adviser, and might be viewed as the precursor to work-focused interviews which now feature in the New Deal programmes, ONE and now Jobcentre Plus. The evaluation of the scheme involved randomly assigning a small group of claimants out of the programme, and comparing their subsequent labour market experiences with those of Restart participants. Although this experimental evaluation method, known as random assignment, is generally viewed as the most reliable method for estimating programme effects, it has its drawbacks, one of which is the ethical issue in denying assistance to claimants which they might benefit from. Another drawback – namely the problems in using survey data from randomly assigned individuals when there has been substantial post-assignment attrition before the survey – resulted in the evaluators using non-experimental methods to analyse the data (White and Lakey, 1992). Since then, very few British evaluations of active labour market programmes have used random assignment but rather have relied on non-experimental data. The adequacy of these techniques was called into question in the United States in the 1980s (for example, LaLonde, 1986). The salience of the random assignment approach in the United States allowed analysts to compare results from experimental and non-experimental techniques using the experimental data to test the bias in non-experimental results. In general, results have not been particularly favourable to non-experimental approaches. However, the research identified circumstances in which particular non-experimental approaches perform well.

Much of the evaluation literature in the United States and Britain now focuses on the value of deploying various non-experimental approaches, and the data requirements that must be satisfied in order to countenance their usage. One of these techniques, known as propensity score matching, is the subject of this paper. Although the technique was developed in the 1980s (Rosenbaum and Rubin, 1983) and has its roots in a conceptual framework which dates back even further (Rubin, 1974), its use in labour market policy evaluation only became established in the late 1990s. It gained particular prominence following the work of Dehijia and Wahba (1998, 1999) who, in reanalysing a sub-set of the data used in LaLonde's (1986) seminal work which had established the superiority of experimental estimators, indicated that propensity score matching performed extremely well. This work has subsequently been criticised in studies which show that propensity score matching, like other non-experimental techniques, depend critically on maintained assumptions about the nature of the process by which participants select into a programme, and the data available to the analyst (Smith and Todd, 2000; Heckman et al., 1998; Heckman and Todd, 2000; Agodini and Dynarski, 2001). Nevertheless, the technique continues to attract

attention as a useful evaluation tool in the absence of random assignment. As we shall see, the method has an intuitive appeal arising from the way it mimics random assignment through the construction of a control group post hoc. Results are readily understood, but less so the assumptions underpinning the validity of the approach and the circumstances in which those assumptions are violated.

The aim of this report is to provide a largely intuitive understanding of the relevance of propensity score matching to evaluation research. As such, it can be seen to be nested within the broader paper by Purdon (2002) which provides an overview of the range of established evaluation techniques. Some overlap with this earlier report is inevitable, however, and in order to achieve a well-rounded report, the assumptions underlying the other techniques will also be presented. This is justified since important extensions to matching include combining it with other techniques. Where appropriate, the issues discussed in this report will be illustrated using the results of recent evaluations, with a particular focus on evaluations in the UK. Several surveys of evaluation techniques already exist. These include Heckman et al. (1999), Blundell and Costa Dias (2000), Smith (2000) and Vella (1998). These sources are all used extensively in the remainder of this report. However, with the exception of Smith (2000) they are all heavy-reading for those not familiar with econometrics. The contribution of this synthesis is to translate the results into a more generally understandable report. This is very timely since the huge academic effort currently being directed to the development of evaluation methodologies, and the particular popularity of matching, means that the literature is extremely fast-moving.

The format for the report is as follows. Section Two describes the evaluation problem and the array of techniques analysts use to tackle it, including matching. Section Three identifies the data requirements for propensity score matching. Section Four, entitled 'Which estimator works?' outlines the advantages and disadvantages of propensity score matching relative to other evaluation techniques. It also identifies the questions that PSM evaluations can answer and those they cannot answer. Section Five describes the stages that the analyst goes through in implementing matching estimators, explaining the difficult choices the evaluator has to make in the course of the analysis.

Section Six gives practical guidance to commissioners of evaluation research as to the circumstances when PSM may be used, and the situations in which it may be appropriate to rule out PSM as an option. To date, few labour market policy evaluations in the UK have been specifically designed for PSM and it is more common for PSM to be used as a method of secondary analysis. One key exception is the evaluation of the national extension of the New Deal for Lone Parents (NDLP) where the wish to apply PSM methods was the driving force behind the evaluation design. The implications of designing for PSM are set out in Section 6.2. The design for the evaluation of NDLP is described in Section 6.3, alongside examples from two other UK programme evaluations. Finally, Section Seven summarises key points from each section.

2. The Evaluation Problem – what it is and how to tackle it

2.1 Why there is a problem

To illustrate ideas, imagine we are interested in measuring the effect of a voluntary training programme on the chances of finding work. At the individual level, we observe the labour market outcomes of those who receive the training and we observe the labour market outcomes of those who do not receive the training. To truly know the effect of the training on a participating individual, we must compare the observed outcome with the outcome that would have resulted had that person not participated in the training programme. However, only one outcome is actually observed. This can be labelled the factual outcome. The so-called counterfactual outcome is that which would have resulted had the participating individual not participated (or the non-participating individual participated). This counterfactual outcome cannot be observed and this is the reason for the evaluation problem. Seen in this way, the essential difficulty of programme evaluation is one of missing data. All approaches to evaluation attempt to provide an estimate of the counterfactual and to use this to identify the programme effect.

It is unlikely that all individuals will respond to a policy intervention in precisely the same way. Rather, there will be heterogeneity in the impact across individuals. This insight raises two questions which evaluations might wish to address. The first is what impact programme participation would have on an individual drawn randomly from the population - the average treatment effect (ATE).[1] The second is what impact participation has on individuals who actually participated – the average effect of treatment on the treated (TT). These two effects are identical if we assume homogeneous responses. However, where we allow for the more realistic scenario of responses varying across individuals, the measures can likewise differ. To illustrate, where a programme is voluntary, we might anticipate that those who volunteer differ from the wider eligible population in terms of their expected gains from the programme: it is because they perceive benefits from participation that they participate in the first place. If this is so, it is unlikely that impact estimates for participants will be relevant for eligible non-participants. Both estimates are of interest. While TT can indicate the average benefit of participation, ATE would be relevant were the policy interest focused on making a voluntary programme compulsory, for example.

It is important for policy-makers to be aware of these different treatment effects for two reasons. First, when comparing results across studies, the reader needs to be aware of which treatment effect the study is addressing. In general, if those with the largest expected gains participate, ATE will be smaller than TT. Second, different policy questions are addressed by the different treatment effects. For example, TT is the parameter which can answer the policy question of whether or not a programme should be abandoned since, if the mean pecuniary impact of treatment on the treated lies below the per-participant cost of the programme, there is a strong case for its elimination. When deciding whether to make a voluntary programme compulsory, thus extending it to the whole eligible population, the question becomes whether or

[1] Note that 'treatment' is the term conventionally used in the evaluation literature to indicate those participating in a programme.

not the mandatory programme would pass a cost-benefit test. (There are difficulties in predicting the impact of a mandatory programme from evaluation of a voluntary one). In this case, the parameter of interest is the ATE. Its ability to account for heterogeneous programme effects is one of the advantages of propensity score matching relative to some other evaluation techniques, as discussed later.

There is a third parameter of interest to policy-makers. This is the impact of a policy introduced to affect people at the margin of participation. The mean effect on those people whose participation changes as a result of the policy is known as the local average treatment effect (LATE). Since the most realistic policy option is often a modest expansion or contraction in the number of persons participating in a programme, LATE may often be the true parameter of interest. However, since it can not be estimated through matching we only refer to it in this paper when discussing other estimation techniques.

2.2 Solutions to the problem?

To understand the problems surrounding programme evaluation it is instructive to consider in more detail our example of a voluntary training programme. A simplistic approach to estimating the programme effect would be to compare the outcome of programme participants with those of non-participants. This would be a valid approach were those participating in the programme a random sample of all those eligible. However, as already noted, this is unlikely to be the case. In reality, such a simple comparison would result in a probable over-estimation of the effectiveness of the programme. If those with more favourable labour market characteristics were more likely to have chosen to participate, it is likely that participating individuals would have done better on average than non-participating individuals irrespective of whether they actually undertook the training. This is the essence of the selection problem. To arrive at a valid estimate of programme impact, the effect of selection must be accounted for.

The effect of the training programme can be characterised by two distinct processes. First, there is the decision to participate in the programme. Second, there is the process determining job entry. There may be numerous influences on job entry, but the influence of interest is that relating to participation in the training programme. That is, the main interest is whether training influences job entry. In these two processes, it is important to realise that both observable and unobservable characteristics may play a role. Examples of the former commonly found in data available to evaluators include age, region and level of qualification. Examples of the unobservables vary according to the quality of the data but, even in the richest datasets, aspects such as motivation, confidence and general ability are rarely measured.

The question of selection bias arises when some component of the participation decision is relevant to the process determining success in job search. More simply, selection bias can result when some of the determinants of participation also influence the outcome. It may be that the relationship between the two processes can be fully accounted for by observable characteristics. In this case, selection bias can be avoided simply by including the relevant variables in the equation explaining outcomes. In the more general case, unobservable characteristics affecting

participation can also influence outcomes. As an example of this, there is much concern in the literature about the impact of individual characteristics such as motivation and orientation to paid employment. It may be that more highly-motivated individuals are more likely to participate in training and are also more likely to find work. In the context of the two process characterisation outlined above, the effects of motivation are correlated across the two equations. Controlling for differences in observable characteristics does nothing to alleviate this. Without addressing the issue of sample selection, the estimated treatment effect will be biased. However, it is worth mentioning that judicious use of observable characteristics can go some way towards minimising the bias associated with unobservables. In the example above, observables which are thought to be highly correlated with motivation, such as pre-programme unemployment duration, may capture some of the motivation effect.

As another example of potentially unobservable differences between the treatment and comparison groups, consider the role of the administrator in selecting participants. Programme entry may be a function of administrator selection as well as choice on the part of the individual applicant. Where administrators are discriminating between the less and better able, either consciously or otherwise, as a basis for programme selection, this process will bias estimates of programme effects if it is unobserved by the evaluator. This may occur where administrators are 'cream-skimming', that is, taking the best for the programme, in which case programme effects will be over-estimated. Equally, programme effects may be underestimated if the programme administrators are targeting programme resources on the least able.

A number of alternative approaches exist that take explicit account of the selection issue. These can be grouped under the broad headings of experimental and non-experimental approaches and are described briefly below. Such an overview is useful to appreciate how the method of matching fits in with other techniques in the evaluator's toolbox. It is also helpful to have this overview because important extensions to matching have included combining the approach with these other techniques.

Before turning to these techniques, it is worth noting that they all share one common feature: they ignore the impact a programme may have on outcomes and behaviour of non-participants. These effects, known as general equilibrium effects, may arise where participants benefit to the detriment of non-participants. This may occur, for instance, where the training participants receive helps them take jobs which would otherwise have gone to non-participants, such that participants simply substitute for non-participants. General equilibrium effects can negate the gains which a partial equilibrium framework suggests accrue to participants. Whether this occurs in practice depends on the nature and size of the programme. For example, a small programme operating in a sizeable labour market is unlikely to generate noticeable general equilibrium effects. Programmes which increase the effective supply of labour by equipping the previously inactive with marketable skills will also have minimal negative effects on non-participants where the programme raises employers' demand for labour. For a detailed discussion of the magnitude and relevance of general equilibrium effects to evaluations using a partial equilibrium framework see Smith (2000).

2.2.1 The experimental approach

Random assignment (or 'experiments') is generally viewed as the most robust evaluation approach (Burtless, 1995). This operates by creating a control group of individuals who are randomly denied access to a programme. Properly carried out, random assignment creates a control group comprising individuals with identical distributions of observable and unobservable characteristics to those in the treatment group (within sampling variation). Hence, the selection problem is overcome because participation is randomly determined. The mean outcome for those participating in the programme relative to that for those in the control group provides an estimate of the TT. While this is the parameter most commonly examined using random assignment, it is possible to design experiments in such a way as to derive estimates of ATE. For example, in the Restart evaluation individuals were randomly excluded from an otherwise mandatory programme (White and Lakey, 1992)

There are, however, a number of provisos. At the practical level, experiments are often costly and require close monitoring to ensure that they are effectively administered. They may also require informing potential participants of the possibility of being denied treatment. The potential for denying treatment can pose ethical questions that are politically sensitive. These may reduce the chances of an experiment being considered as a means of evaluating a programme and may also increase the chances of those responsible for delivery of the programme being reluctant to cooperate. In the UK, there are very few examples of experiments. An early exception was the evaluation of Restart (White and Lakey, 1992). More recently, the New Deal for the Long-term Unemployed (NDLTU) included an experimental component. In two of the twenty-eight NDLTU pilot areas the treatment was randomly allocated on the basis of odd or even National Insurance numbers.

There are also practical problems that can bias the estimates. It may be that the implementation of the experiment itself alters the framework within which the programme operates. This is known as 'randomisation bias' and can arise for a number of reasons (Heckman and Smith, 1995). For instance, if random exclusion from a programme demotivates those who have been randomised out, they may perform more poorly than they might otherwise have done, thus artificially boosting the apparent advantages of participation. Furthermore, those receiving treatment may drop out of the programme. In this case, random assignment does not identify treatment on the treated but instead identifies the mean effect of 'intent to treat'. This may or may not be of direct policy interest. Conversely, those denied treatment may choose to participate in programmes that are effective substitutes for the programme under evaluation. With both programme dropout and comparison group substitution, non-experimental methods can be used to retrieve the desired parameters. However, this is a second-best position since experiments are designed specifically to avoid this sort of adjustment. Moreover, it is worth noting that the problems of programme drop-out and comparison group substitution are not unique to experiments, although experiments may exacerbate the second problem by creating a pool of people who want to participate but were refused by the randomisation process.

2.2.2 Non-experimental approaches

Despite the potential drawbacks associated with random assignment, correctly administered, it remains by far the most robust means of estimating programme effects. However, many programme evaluations are carried out using non-experimental approaches. There are a number of non-experimental evaluation techniques and the choice of best approach is determined in large part by practicalities. Specifically, the characteristics of the programme and the nature and quality of available data are key factors. These non-experimental techniques all share one thing in common: in the absence of an observable counterfactual, assumptions have to be made to identify the causal effect of a policy or programme on the outcome of interest. We call these 'identifying assumptions'. In general, the fewer assumptions you make, and the more plausible they are, the more likely it is that estimated effects will approximate real programme effects. In this section, the main approaches are discussed and their identifying assumptions are highlighted. They are presented under two broad categories: before-after estimators and cross-section estimators.

Before-after estimators

The essential idea of the before-after estimator is to compare the outcomes of a group of individuals after participating in a programme with outcomes of the same or a broadly equivalent group before participating and to view the difference as the estimate of TT. This approach has been widely-used in evaluations, usually adjusting the results to control for the effect of observable characteristics.

Before-after estimators concern themselves with selection on unobservables. The identifying assumption for this estimator is that the difference between the true post-programme counterfactual and their pre-programme outcome averages out to zero across all individuals participating in the programme. In fact, so long as this averaging out takes place, the approach does not require longitudinal data, but can be implemented instead with repeat cross-section data so long as at least one cross-section is from a pre-programme period. In essence, the before-after estimator assumes that the unobservables are of two types: those specific to an individual and fixed over time (individual effects) and those specific to an individual but not fixed over time (transitory effects). Participation in the programme is held to depend on the fixed effect and not the transitory effect. This is a strong assumption. Clearly, macroeconomic changes between the two observation points will violate the assumption as might changes in the life-cycle position of a cohort of participants.

In view of the likely transgression of the identifying assumption, a more widely-used approach is the difference-in-differences (DiD) estimator, also known as the 'natural-experiment' approach (see, for example, Blundell and MaCurdy, 1999). This operates by comparing a before-after estimate for participants with a before-after estimate for non-participants and regarding the difference as TT. The identifying assumption is more plausible than that for the before-after estimator. Specifically, the average change in the no-programme outcome measure is assumed to be the same for participants and non-participants. What this means in effect is that the DiD estimator can cope with macroeconomic changes or changes in the lifecyle position so long as

those changes affect both participants and non-participants similarly. This highlights the need to select a suitable comparison group of non-participants. Often, the choice of comparison group is justified on the basis of it trending in a similar way to the treatment group with regard to the outcome variable in question over a prolonged period of time preceding the introduction of the programme. While this is reassuring, it is worth bearing in mind that it is usual to adjust DiD estimates for observable characteristics and consequently it is the regression-adjusted outcomes that should trend together rather than the outcome measures themselves.

The effectiveness of the DiD estimator can be seen by considering the nature of the characteristics of the unobserved variables that may affect outcomes. In addition to the individual effects and transitory effects characterising the before-after estimator, an effect common to individuals but varying over time (trend effect) is also allowed for. As already noted, the before-after estimator eliminates the individual effects. The advantage of the DiD estimator is that it also removes the trend effects. Thus, the only remaining effect is that specific to the individual but varying over time. This cannot be controlled for, and should it influence the decision to participate in the programme the identifying assumption will be violated and the resulting estimates biased.

The fragility of the DiD estimator to violation of the identifying assumption can be seen by considering an empirical phenomenon which has become known as Ashenfelter's dip. It has often been noted that participation in a training programme is more likely where there is a temporary reduction in earnings just before the programme takes place. Should earnings be mean-reverting, growth among participants would exceed that among non-participants irrespective of whether they received any training. Consequently, the DiD estimator (and the before-after estimator) will provide over-estimates of programme effects on earnings in this scenario (Heckman and Smith, 1999).

Furthermore, the before-after estimator and the DiD estimator both rely on the composition of the treatment group remaining unchanged in the post-programme period. If this is not satisfied, the difference between the true counterfactual and the pre-programme outcome will not necessarily average out to zero across all individuals. Changing composition can occur most obviously with repeat cross-section data but it is also possible with longitudinal data should the sample deplete over time on a systematic basis.

Cross-section estimators

If longitudinal or repeat cross-section data are not available, other approaches must be considered. Cross-section estimators use non-participants to derive the counterfactual for participants. Until recently, the standard way in which to isolate the independent effect of programme participation on labour market outcomes would have involved controlling for observable differences between participants and non-participants using regression methods. For example, to compare differences in the rates of a binary outcome (such as entry to work) between participants and non-participants after controlling for observable differences in the two groups, it would have been standard practice to use a logistic regression approach with the binary outcome as the dependent variable, and the observables plus a binary 'participation' indicator, as the

independent variables. The coefficient for the 'participation' indicator would be interpreted as the programme effect on the treated (i.e. TT) after controlling for the observables. Implicit in this approach is the assumption that, having controlled for observables, participation is independent of the process determining outcomes. In other words, observables which enter the regression capture selection into the programme. As we note below, regression shares this assumption with the method of matching. First, we discuss two other cross-sectional estimators common in the literature which deal with selection on unobservables. These are instrumental variables and the Heckman selection estimator.

Instrumental variables

The IV method is possible when a variable can be identified that is related to participation but not outcomes. This variable is known as the 'instrument' and it introduces an element of randomness into the assignment which approximates the effect of an experiment. Where it exists, estimation of the treatment effect can proceed using a standard instrumental variables approach.

Where variation in the impact of treatment across people is not correlated with the instrument, the IV approach recovers an estimate of impact of treatment on the treated (TT). However, if the variation in gains is related to the instrument, the parameter estimated is LATE (Imbens and Angrist, 1994). Consider the example of using distance from a training centre as the instrument. If individuals know their gains from treatment, then among participants, those from farther away need a larger gain than average to cover their higher cost of participating. Where there is such a correlation, LATE is estimated. As noted earlier, if the policy under consideration is a marginal increase or decrease in the costs of participation, then LATE is the parameter of interest.

The main drawback to the IV approach is that it will often be difficult to find a suitable instrument because, to identify the treatment effect, one needs at least one regressor which determines programme participation but is not itself determined by the factors which affect outcomes (Blundell and Costa Dias, 2000; Heckman, 1995).

Heckman selection estimator

This approach has been used extensively in evaluations, especially in the UK.[2] It allows for selection into the treatment group on the basis of variables that are unobservable to the analyst. It operates by assuming a particular form for the distribution of the unobservable characteristics that jointly influence participation and outcome.[3] By explicitly modelling the participation decision, it is possible to derive a variable that can be used to control for that part of the unobserved variation in the outcome equation that is correlated with the unobserved variation in the participation decision. Including this new variable alongside the observable variables in the outcome equation can result in unbiased estimates of the treatment effect. While not strictly necessary from a mathematical viewpoint, credible implementations include

[2] In its common form, it is a specialisation of the control function approach (Heckman and Robb, 1985, 1986)

[3] More technically, the errors in the participation and outcome equations are usually assumed to follow a bivariate normal distribution.

an instrument; that is, a variable included in the estimation of the participation equation that is excluded from the outcome equation.

This approach appears to offer an elegant means of obtaining an estimate of TT in the presence of selection. However, there are two main drawbacks. First, as with the IV approach, the identification of a suitable instrument is often a significant practical obstacle to successful implementation. Second, the resulting estimates are entirely contingent on the underlying distributional assumption relating to the unobserved variables. In fact, research has shown that estimates can be surprisingly sensitive to these assumptions not being met.[4]

Method of matching

The method of matching has achieved popularity more recently as a tool of evaluation. It assumes that selection can be explained purely in terms of observable characteristics. Applying the method is, in principle, simple. For every individual in the treatment group a matching individual is found from among the non-treatment group.[5] The choice of match is dictated by observable characteristics. What is required is to match each treatment group individual with an individual sharing similar characteristics. The mean effect of treatment can then be calculated as the average difference in outcomes between the treated and non-treated.

The approach has an intuitive appeal but rests on two assumptions. The first is that if one can control for observable differences in characteristics between the treated and non-treated group, the outcome that would result in the absence of treatment is the same in both cases. This identifying assumption for matching, which is also the identifying assumption for the simple regression estimator, is known as the Conditional Independence Assumption (CIA). It allows the counterfactual outcome for the treatment group to be inferred, and therefore for any differences between the treated and non-treated to be attributed to the effect of the programme. To be credible, a very rich dataset is required since the evaluator needs to be confident that all the variables affecting both participation and outcome are observed. That is, it is assumed that any selection on unobservables is trivial in that these unobservables do not affect outcomes in the absence of the treatment. Where data do not contain all the variables influencing both participation and the outcome, CIA is violated since the programme effect will be accounted for in part by information which is not available to the evaluator. One example might be instances in which the evaluator is unaware of those approaching eligibility for the programme adjusting their behaviour in anticipation of programme entry by reducing their job search (the Ashenfelter dip discussed earlier). This might affect both their probability of entering the programme and their likelihood of obtaining a job. Another example is where some of those eligible for a programme do not participate because they are expecting a job offer shortly. However, if the CIA holds, the matching process is analogous to creating an experimental dataset in that, conditional on observed characteristics, the selection process is random. Consequently, the distribution of the counterfactual outcome for the treated is the same as the observed outcomes for the non-treated.

[4] See, for example, Goldberger (1983) and Puhani (2000)

[5] In practice, participants may be matched to multiple non-participants. We return to this later in the paper when various types of matching are considered.

Matching makes an assumption made by all partial equilibrium estimators, namely that an individual's programme participation decision does not depend on the decisions of others. This assumption would be violated if peer effects influenced participation. This might occur, for instance, where a programme targeted at lone parent benefit claimants is highly regarded by participants locally, thus encouraging others to join. If this peer correlation is unrelated to outcomes it becomes part of the error term in the participation equation and need not be a problem. However, where peer correlation is related to outcomes, estimates that can not account for those peer effects will be biased. An example might be instances in which a programmes offers a limited number of places such that decisions to participate now reduce the probability of other applicants entering the programme later. If decisions to participate early are correlated with factors which independently improve labour market prospects – such as motivation to get a job – estimates failing to account for this will be upwardly biased. It is possible to overcome this problem where proxies for supply constraints are available for inclusion in the estimation (Sianesi, 2001).

Another assumption that is required for matching and all of the other partial equilibrium estimation strategies is the so-called SUTVA assumption, where SUTVA means stable unit treatment value assumption. This assumption says that the impact of the programme on one person does not depend on whom else, or on how many others, are in the programme. As Sianesi notes (2001) SUTVA is in fact the assumption that the model's representation of outcomes is adequate, that is that the observed outcome for an individual exposed to treatment depends only on the individual and not on what treatments other individuals receive nor on the mechanism assigning treatment to individuals and that whether the individual participates only depends on the individual.

A practical constraint exists in that, as the number of characteristics used in the match increases, the chances of finding a match reduce. It is easy to see that including even a relatively small number of characteristics can quickly result in some participants remaining unmatched. This obstacle was overcome thanks to an important result (Rosenbaum and Rubin, 1983) showing that matching on a single index reflecting the probability of participation could achieve consistent estimates of the treatment effect in the same way as matching on all covariates. This index is the propensity score[6] and this variant of matching is termed 'propensity score matching'. Its clear advantage is that it replaces high-dimensional matches with single index matches.

This problem of reduced chances of finding a match does not disappear entirely with propensity score matching, however. It is still possible that there will be nobody in the non-treatment group with a propensity score that is 'similar' to that of a particular treatment group individual.[7] This is known as the 'support' problem and means of addressing it exist that vary in their level of complexity. However, they all operate by identifying participants who are poorly-matched and then omitting them from the estimation of treatment effects. What they seek to guarantee is that any combination

[6] Note that the index used for matching need not necessarily be the probability of participation, although in practice it commonly is. More consideration will be given to this later in the report.

[7] Participants are similar to non-participants in that their propensity to participate is similar. For this to happen, it is not necessary for the participant and matched non-participant to share characteristics. Rather, the values each has for the combination of variables entering the participation equation generate similar propensity scores.

of characteristics seen among those in the treatment group may also be observed among those in the non-treatment group.

Where there is no support for the treated individual in the non-treated population, this treated individual is dropped from the analysis. The estimated treatment effect has then to be redefined as the mean treatment effect for those treated falling within the common support. In one way, this is a strength of matching, since it makes explicit the need for common support across the treated and non-treated. However, enforcement of the common support can result in the loss of a sizeable proportion of the treated population, especially when considering multiple-treatment programmes For example, Gerfin & Lechner (2000) reject 14 per cent of their sample, while Frölich et al. (2000) reject 27 per cent. One must bear this in mind when considering the policy relevance of results. This is because the policy analyst wishes to know the effect of a policy on those who participate (or even the whole eligible population), not just a sub-sample for whom common support is enforceable. If treatment effects differ non-randomly with those unsupported characteristics, the treatment effect relevant to the supported sub-population will not provide a consistent estimate for the unsupported sub-population. Whether this is a problem in practice will depend upon the proportion of the treatment group lost. In any event, it is informative to consider the characteristics of those treated lost to the analysis since this will uncover the sorts of treated individuals who have no counterparts in the non-treated population. This can often tell us a great deal about the nature of selection into a programme and provide important clues for the interpretation of estimated effects.

The explicit acknowledgement of the common support problem is one of the main features distinguishing matching methods from standard parametric regressions. The problem is less severe with parametric approaches since the model results can be used to extrapolate to unsupported regions. However, such out-of-sample predictions will need to be carefully assessed. The other main distinguishing feature, already hinted at, is that matching is non-parametric. Consequently, it avoids the restrictions involved in models that require the relationship between characteristics and outcomes to be specified. If one is willing to impose a linear functional form, the matching estimator and the regression-based approach share the same identifying assumptions.

While the use of matching has largely focused on participation or non-participation in a programme, there is often a need to consider programmes which may comprise different types of treatment. This is particularly true when considering European labour market programmes. Imbens (1999) and Lechner (2001a) show that the major results relating to the two-state framework extend straightforwardly to the case of multiple mutually-exclusive treatments. Hence, matching can be used to evaluate more complex labour market programmes.

3. Data Requirements for Propensity Score Matching

Returning to PSM, it is clear from the discussion above that the assumptions underpinning both PSM and regression methods rely on adequate data on matters affecting participation and outcomes. In this section, we consider some of the data dilemmas facing evaluators, and suggest some partial solutions.

How does the evaluator know which 'observables' affect both programme participation and the outcome of interest, and thus merit inclusion in the matching process? There is no clear answer to this question: the evaluator is guided by economic theory and prior research (which might include specially commissioned research on the reasons for participation). The CIA is an untestable assumption, so the evaluator will never know whether it has been met.[8] Knowledge of the empirical and theoretical literature is thus a prerequisite for designing a survey intended for analysis using PSM, or when faced with the need to choose a sub-set of variables from a rich data set that has already been collected. Fortunately, there is an abundance of theory and evidence relating programme participation to labour market outcomes which facilitates the evaluator's choice of variables. However, in other areas, this information is lacking. For instance, although matching might offer an attractive means for evaluating the impact of tax credits on employer costs, recruitment and wage policies, it is by no means clear what variables would be appropriate for inclusion in matching.

In the sphere of labour market evaluation, there is substantial evidence to indicate that unemployment and earnings histories prior to programme entry are key. The most comprehensive review of training programme evaluations for the United States indicates that the wide variation in estimated training programme effects from non-experimental evaluations stems from 'a lack of sufficient information to adjust fully for pre-program differences between participants and non-participants' (Friedlander, Greenberg and Robins, 1997: 1832-1833). As a result, the absence of good pre-programme earnings data has hampered estimates of training effects on post-programme earnings (Friedlander and Robins, 1995). Pre-programme unemployment and earnings are valuable variables when estimating with PSM because they are important predictors of programme entry and subsequent employment outcomes in their own right. They can also help capture otherwise unobservable characteristics, such as motivation, which might also influence participation and outcomes.

Even if strong theory exists to indicate the sorts of variables the evaluator is going to need, and these variables are observable in theory, one still needs the time and resources to collect those data, and to do so with minimal measurement error. If the data set required is large, survey costs may prove prohibitive. In these cases, the evaluator will need to prioritise those variables that, according to theory and evidence, are most likely to influence participation and outcomes. A further constraint arises when minimal pre-programme data are available for important variables. This can be the case, for instance, for attitudinal or motivational data relevant to both participation and outcomes. It is difficult for these data to be retrospectively calculated or recalled

[8] If the analyst is willing to assume that the participation and outcome processes are the same across time one can test the CIA, for example, by using pre-programme tests such as those outlined by Heckman and Hotz (1989).

by respondents. In a study of careers guidance for employed adults, White and Kileen (2000) show PSM estimates of treatment effects are highly sensitive to the use of attitudinal data collected shortly after programme entry. Considerations such as these were the motivating force in the design of the evaluation of the national extension of NDLP (see section 6.3). Data on administrators' assessments of applicants are also important in understanding selection processes, but they are often absent. Where these data and administrators' records of offers made to applicants to join programmes are available to the evaluator they have markedly improved the quality of the matching (Sianesi, 2001).

If key data are missing for a sub-set of observations this can substantially affect the estimation of treatment effects using non-experimental estimators such as PSM. Thus, the decision to select a sub-sample for whom pre-programme earnings were available seems to explain why two classic studies which used otherwise identical data came to different conclusions about the value of non-experimental estimators (the studies in question are LaLonde (1986) and Dehija and Wahba (1999) – see Zhao (2000) and Smith and Todd (2000) for evaluations of these studies). The evaluator is in a no-win situation in these circumstances: from a PSM perspective, the CIA is not met when a variable likely to predict outcomes and participation is overlooked, while the decision to select a sample for whom key data are available has the potential to produce quite different results to those one might have expected were the data available for the full sample (Smith and Todd, 2000). If the evaluator wishes to estimate participation with the fuller sample, she can try to take account of missing data by imputing values (for example, using regression techniques) and incorporating dummy variables identifying cases with imputed data (see, for instance, Smith and Todd, 2000: 19, footnote 34).

Propensity score matching is 'data hungry' not only in terms of the number of variables required to estimate participation and outcomes, but also in the number of participants and non-participants entering the matching process. One needs to be wary when interpreting results based on small samples of non-participants. This is because, for PSM to be tenable, as opposed to matching on all covariates, treatment and comparison group members with similar propensity scores should be balanced along the covariates (the Xs) entering the matching process which affect outcomes and participation. One must allow for the possibility of neighbourhood matching where exact matches are unlikely, as in the case where the matching criteria include continuous variables. When matching is exact at the propensity score, then the distribution of *X*s which predict participation and outcomes will be the same for the treated and comparison samples. However, where matching occurs on some neighbourhood of the propensity score, as is usually the case, the distribution of X is only approximately the same for the treated and non-treated sample within the neighbourhood of that propensity score. Provided there is a sufficient number of observations on participants and non-participants across values of X, this balance between participants and non-participants can be achieved by matching on the propensity score. But for this balancing property to hold, a decent sample size at each value of the propensity score is needed (Zhao, 2000). The averaging out of mismatches, which relies on the balancing property, therefore requires a large number of observations. In practice, empirical studies are often based on relatively small samples, potentially biasing results or making them subject to large sampling errors. This consideration leads Zhao (2000) to estimate the performance of matching

estimators other than a pure propensity score using different sample sizes and data sets. He concludes that the performance of different matching estimators varies case-by-case and depends largely on the data structure at hand.

More generally, smaller samples will provide less precise estimates of treatment effects and may result in small effects passing undetected. This is the case with other methods. However, a complication that arises in the case of matching is that the estimation of the propensity score and the matching itself both add variation beyond normal sampling variation (Heckman, Ichimura and Todd, 1998). This must be accounted for when calculating the standard errors of the estimates. The number of participants and non-participants required for a PSM evaluation will vary dependent upon the nature of the programme and the eligible population. In general, smaller samples will be needed where a single-treatment programme is being evaluated than in circumstances where a multiple treatment programme is being evaluated because the support problem is more onerous in the multiple treatment case. It is worth noting at this point that the complexities of PSM mean that to calculate the standard errors of estimates requires computer-intensive replication methods such as bootstrapping. [9]

In Britain as elsewhere, programme administrators have increasingly recognised the value of administrative data, collected as part of the programme participation process and often stored as programme eligibility databases, in evaluating programme impacts. It is clear that databases such as the one for the New Deal for Young People are sufficiently rich in data predicting outcomes and participation that they can support persuasive evaluations based on PSM (Dorsett, 2001). These data have the greatest potential where they contain similar information on participants and non-participants within the eligible population. The value of these data is that they contain a) pre-programme unemployment records which, as noted above, are critical in estimating participation and post-programme outcomes b) fixed individual characteristics c) time-coded programme interventions, permitting a more accurate picture of treatment sequences (which some studies find have a profound influence on outcomes (Sianesi, 2001)) and accurate data on time to programme entry and d) time-coded data on post-programme experiences of unemployment which can be used to derive accurate estimates of post-programme effects. Their drawback is that they do not generally include data on attitudes and motivation, which may be important when evaluating particular programmes.

[9] Empirical investigation indicates that above a certain sample size bootstrapped standard errors are roughly equivalent to those which ignore the additional variation in PSM. However, this sample size is not trivial (Heckman, Ichimura and Todd, 1998).

4. Which Estimator Works?

Clearly, there is no single answer to this question. The choice of evaluation technique depends on the nature of the programme being evaluated plus the nature of the available dataset. The identifying assumptions that underlie each of the approaches may be more or less credible in particular applications. The key point is that all approaches involve assumptions being made, and these assumptions are generally untestable in practice. However, researchers have sought to appraise the plausibility of key assumptions by analysing experimental data using non-experimental techniques, judging the 'success' of the non-experimental approaches by their ability to replicate experimental results. Findings are mixed, with PSM techniques successfully replicating experimental results in one study (Dehijia and Wahba, 1999) but not in others (Heckman et al., 1998; Heckman and Todd, 2000; Agodini and Dynarski, 2001). Heckman at al. (1998) is the most comprehensive attempt to use experimental data to examine the assumptions underlying various evaluation techniques. While the results are data-dependent and not necessarily generalisable, they provide an insight into the strength of the assumptions underlying the main techniques. Importantly, the assumption underpinning matching is rejected. However, Heckman et al. also consider combining matching with DiD. This 'conditional' DiD estimator allows for individual fixed effects and trend effects to influence participation. In other words, it weakens the identifying assumption for matching by allowing unobserved variables to influence participation. This weaker assumption was not rejected in their study[10]. In general, studies comparing experimental and non-experimental results place non-experimental estimators at a disadvantage because they rely on drawing comparison groups from localities other than those where the treatment occurs, and often use different data to construct predictor and outcome measures (see, for instance, LaLonde, 1986; Dehijia and Wahba, 1999; Smith and Todd, 2000; Agodini and Dynardski, 2001).[11]

Even if, in a particular study, one or other of the assumptions underpinning matching is likely to be violated, this does not mean that we should dismiss using PSM out of hand. It is also important to consider the likely seriousness of the violation, and the direction of any bias introduced. For instance, there is evidence that controlling for bias due to observable characteristics is more important than controlling for the bias due to unobservables (Heckman et al., 1998). Even if the assumptions are not justifiable and there is no prior knowledge of the relative magnitude of the bias due to unobservables and the bias due to observables, it is useful to apply matching methods to eliminate the bias due to observables first and then use different procedures to address the bias due to unobservables, such as the use of the conditional DiD estimator described above.

[10] For an application of conditional DiD to the evaluation of New Deal see Blundell et al. (2001) and Blundell and Costa Dias (2000). See Bergemann, Fizenberger and Speckesser (2001) and Eichler and Lechner (2000) for applications of conditional DiD together with matching.

[11] The comparisons using JTPA data in Heckman et al. (1998) are different in this respect because JTPA eligible non-participants are from the same local labour markets and have data collected in the same manner as a subset of people in the experiment.

4.1 Advantages and disadvantages of PSM relative to other evaluation techniques

PSM has two clear disadvantages relative to experimental techniques. The first is the need to make the CIA. In the case of random assignment, properly conducted, we can be confident that the treated and non-treated populations are similar on both observable and unobservable characteristics. This is not true in the case of PSM, which takes account of selection on observables only. Second, whereas PSM can only estimate treatment effects where there is support for the treated individuals among the non-treated population, random assignment ensures that there is common support across the whole sample. These considerations make experimental techniques unambiguously superior to PSM. However, practical considerations are also important in the design and execution of programme evaluations and, in some circumstances, these practical considerations may favour PSM over random assignment.

PSM's main advantage over random assignment is that it avoids the ethical considerations which arise when a potentially beneficial treatment is denied to those randomly assigned out. Cost is also an important practical consideration when conducting evaluations. In some cases, despite PSM's onerous data requirements, data generation may be less costly than in the case of an experiment since the latter involves substantial monitoring to secure the random allocation. That said, if outcome data are available through administrative sources, experimental evaluations can often do entirely without surveys, keeping costs low.

Sometimes, the advantages that random assignment offers in principle can diminish in practice due to data collection problems. Key among these is sample attrition. Some of the advantages of random assignment are lost when there is substantial sample attrition post-randomisation. In these circumstances, as in non-experimental evaluation, one must adjust for sample attrition by incorporating suitable weights. Substantial loss of individuals from the data can compromise any attempt to estimate programme effects, whatever the method used (see Sianesi, 2001 for an account of this problem in a PSM evaluation). Another difficulty encountered in experimental evaluation is the possibility that programme effects may be misleading if those randomly assigned out of the treatment have other, perhaps similar, training or job search opportunities available to them. In these circumstances, 'non-treatment' equates to treatment with something else, not the absence of treatment. This is significant from a policy perspective, because policy customers may equate the impact of treatment on the treated with the impact of treatment versus having no treatment at all. Even if survey information provides accurate information as to the alternative programmes entered by the control group, experimental methods can only estimate the impact of the randomly assigned treatment relative to not having that treatment. Where a proportion of the control group have accessed a programme similar to the one they were randomly assigned out of, the evaluator must resort to non-experimental methods, or adjustments in the experimental data to account for substitution. With PSM, the evaluator can designate the nature of the treatment, and so is not bound by the need to focus purely on the treatment that may have been randomly assigned.

What of the advantages of PSM relative to other non-experimental evaluation techniques? Matching is unambiguously preferred to standard regression methods for two reasons. First, matching estimators highlight the problem of common support, since treatment effects can only be estimated within the common support. Where there is poor overlap in support between the treated and the non-treated this raises questions about the robustness of traditional methods relying on functional form to extrapolate outside the common support. Secondly, matching does not require functional form assumptions for the outcome equation (that is, it is non-parametric). Regression methods impose a form on relationships (usually linear) which may or may not be accurate and which PSM avoids: this is valuable since these functional form restrictions are usually justified neither by economic theory nor the data used (Dehejia and Wahba, 1998; Smith and Todd, 2000). However, if there are strong indications about the nature of the functional form for the outcome equation for the population in question, either from theory or from earlier empirical research, there are efficiency gains from imposing it in the estimation.

Whether matching is advantageous relative to methods that deal with selection on unobservables depends on the available data and on the institutional nature of the selection process into the treatment. Matching is only feasible where there is a firm understanding, based on theory and past empirical evidence, of the determinants of programme participation and the outcomes of interest. If this information is available, and the data are available to make the CIA plausible, then matching is feasible. This avoids the search for a good instrument by which to identify the selection process and separate this process from the one governing outcomes. The appropriateness of the exclusion restriction identifying the two equations, required for the IV approach and the Heckman selection estimator described earlier, is an untestable assumption, and one which is often inherently difficult to make. This is because it is often difficult to find variables affecting the probability of programme participation which do not affect employment outcomes other than through their effect on participation. However, if such instruments are available, the IV and Heckman selection methods are feasible. If there are no good instruments available and the CIA is not plausible, but longitudinal outcome data are available and selection is plausibly on the fixed component of the unobservables, the analyst can use the difference-in-difference estimator, perhaps in combination with matching.

Furthermore, with matching, there is no need for the assumption of constant additive treatment effects across individuals which is required in simple regression and Heckman and bivariate normal selection estimation procedures. Instead, heterogeneous treatment effects are permitted, and can be retrieved via sub-group analysis. This involves selecting the sub-group of interest and re-matching within that group. This makes PSM a flexible tool for studying programme effects on groups of particular interest.

4.2 The questions that PSM evaluations can answer and those they cannot answer

It is important to bear in mind what PSM can do, but also to note its limitations as an evaluation tool. Its strength lies in estimating mean programme effects for a population or sub-group. In doing so, it can estimate treatment on the treated . Where

the outcome is earnings, the mean effect of treatment on the treated can be the basis for a cost-benefit analysis which may determine the viability of a programme. Policy makers may view a programme as viable if the estimated benefit for the treated outweighs the per-participant cost. It can also estimate the average treatment effect, namely the mean impact on all persons eligible for treatment, not just those who voluntarily participate. This is useful for estimating the possible impact of making a voluntary programme compulsory, or extending the programme beyond the current eligible group. To obtain ATE using PSM the treated have to be used to estimate the counterfactual for the untreated. This means that, when enforcing common support, it is not sufficient to have support for the treated among the non-treated. One also needs support for the non-treated in the treated population. This does mean that, where there is only a small pool of participants from which to draw comparators for the non-treated, there may be a support problem which constrains the evaluator's ability to estimate the average effect. Once common support is enforced, TT and treatment on the untreated have to be estimated, with the ATE being a weighted average of the two.

There are three questions which PSM cannot address, but which may be of crucial importance to the policy maker. First, in common with the experimental and all the non-experimental techniques discussed in this paper, PSM cannot answer questions relating to the distributional effects of the programme, such as the percentage of programme participants who benefit. This is because none of these techniques allows for reliable assumptions to be made about the distribution of impacts (Heckman, Smith and Clements, 1997). PSM can only recover mean effects. Second, as is the case with other partial equilibrium evaluation techniques, it can not establish the impact of the programme beyond the eligible group. General equilibrium effects may bias partial equilibrium estimators such as random assignment and PSM which assume that the outcomes observed for eligible non-participants in the comparison or control group are unaffected by the existence of the programme. In practice, these effects may occur where helping programme participants proves to be at the expense of other disadvantaged groups beyond the programme. These effects, which may result in participants displacing non-participants in paid work, or employers substituting non-participants for participants, may reduce the overall benefits of the programme in the economy at large relative to a picture drawn solely from information about participants. Whether any particular programme will influence outcomes for people beyond the eligible group depends on the size of the programme, how effective it is in benefiting participants, and the extent to which participants are capable of taking the jobs of non-participants.

Thirdly, PSM can not estimate the local average treatment effect, which is the mean impact of the programme on those whose participation status changes due to a change in policy. So it is less useful than IV techniques when considering broadening (or narrowing) eligibility for a programme in a targeted way.

In common with most other quantitative evaluation techniques, PSM is not particularly well-suited to 'getting inside the black box' of how a programme does or does not work. That said, it is capable of revealing dimensions along which selection into the programme has occurred (through the participation equation and enforcing common support), and it can generate programme effects for sub-groups which can also indirectly cast light on the way the programme operates.

Finally, and again, in keeping with other evaluation methodologies, PSM is not capable of giving a 'once and for all' definitive indication of a programme's effects. That will depend upon ambient labour market conditions, the stage in the business cycle, cohort effects (how early/late in the history of the programme, the size of that cohort, its composition – stock versus flow etc), whether the programme is voluntary or compulsory, the size of the programme, and so on.

4.3 Use of PSM under different programme implementation models

PSM can be used to estimate programme effects whenever the programme implementation generates pools of treated and untreated individuals from which the two matched groups can be drawn. For voluntary programmes the two pools are usually taken to be the participants and non-participants respectively. For mandatory programmes the identification of an appropriate untreated pool is less obvious.

For mandatory programmes that are introduced in a few pilot areas before being introduced more widely the untreated pool will usually be the eligible population from non-pilot areas. A concern in this instance will be that the propensity score modelling adequately captures area differences in participation and outcomes so that between-area effects are controlled for.

For mandatory programmes that are introduced nationally without piloting the only untreated pool from which the comparison sample might be drawn is the eligible population from the period *before* the programme was implemented, to give a before-after design. In this instance PSM will control for differences in the profiles of the two eligible populations (before and after) but will not automatically allow for programme effects to be separated out from temporal effects. An additional problem is that eligible persons from the pre-programme period often gain access to the programme once it is in place, making it difficult to obtain clean impact estimates for more than the short term.

In general, observables entered into a propensity score model for a mandatory programme are likely to be very different from those for a voluntary programme since factors such as personal motivation and attitudes towards work will not affect participation in a mandatory programme. If take-up of a mandatory programme is almost total, selection is not an issue: all who are eligible participate. In reality, programmes are only mandatory to the extent that claimants lose benefit entitlements for non-participation. It is often the case that a large minority of the eligible group drop out or refuse to participate, presumably because they view non-participation as the more attractive option. In these cases, selection is an issue but the selection process is somewhat different from the one governing entry to a voluntary programme.

It is worth noting that PSM methods are not restricted to situations where one wants to compare participants with non-participants. PSM can also be used to compare groups who experience different levels or types of participation. In this instance the propensity score would equal the probability of participation at a particular level or of a particular type, and the matched sample would be selected from those experiencing the other, or counterfactual, level or type. Thus, for instance, in a randomised trial

setting, PSM might be used to estimate the impact of different types of 'treatment' amongst those randomised to the treatment arm. In these cases, one needs to be aware of the treatment parameters being identified – it's not the effect of treatment relative to non-treatment but the effect of treatment relative to another type of treatment.

5. Implementing Matching Estimators

While the idea behind matching is simple, much care is needed when constructing a matching estimator. In this section, a number of practical issues are considered. The discussion encompasses aspects of the approach that have received considerable attention in the literature as well as some which are less commonly addressed.

The first and most important point to stress is that matching is not appropriate in all circumstances. In particular, as already noted, matching makes very heavy demands on the data. Only if all variables thought to influence both participation and treatment are observable is the CIA credible. Consequently, matching will only be a suitable candidate as an approach to evaluation when very informative data are available (a condition applying equally to simple regresssion). As a more general point, the issue of data quality should not be overlooked irrespective of which evaluation technique is being considered. Heckman et al. (1998) considered this issue and showed that poor data can be responsible for introducing at least as much bias as can the choice of evaluation technique.

5.1 Estimating programme participation

Since the propensity to participate is unknown, the first task in matching is to estimate this propensity. Any resulting estimates of programme effect rest on the quality of the participation estimate. This can be routinely carried out using a choice model. Which choice model is appropriate depends on the nature of the programme being evaluated. If the programme offers a single treatment, the propensity score can be estimated in a standard way using, for example, a probit or logit model, where the dependent variable is 'participation' and the independent variables are the factors thought to influence participation and outcome. If the programme offers multiple treatments, a model that can represent multi-way choices should be chosen. This might be a multinomial logit or a multinomial probit, for example. In choosing a model, the analyst must pay attention to their acknowledged properties. In the single treatment case, the probit and logit models usually yield similar results so the choice between them may not be critical. In the multiple treatment case, however, it is well-acknowledged that the multinomial logit model is based on stronger assumptions than the multinomial probit model. However, the latter is computationally much more burdensome and it is only relatively recently that it has become possible to use such models in cases when a choice of more than three options is required. Consequently, the multinomial logit model still finds some use (Larsson, 2000, for example).
A compromise solution that offers flexibility while still being simple to estimate is to estimate a series of binomial models, considering participation in each option relative to every other option. This was the procedure adopted by Bonjour et. al. (2001). It is computationally uncomplicated but suffers from two drawbacks. First, as the number of options increases, the number of models to be estimated increases disproportionately.[12] This can result in the researcher being deluged with output. Second, each model considers two options at a time and estimates the probability of choosing a given option. Consequently, this choice is conditional on being in one of the two selected options. As will be seen later, the support requirement mentioned earlier is not defined on such conditional probabilities. This may limit the ability of

[12] Specifically, for M options there will be M(M-1)/2 models to estimate.

the approach to estimate certain programme effects such as the ATE. Lechner (2001b) considers the relative performance of the multinomial probit approach and the approach based on a series of probits in the case of a Swiss programme. He finds little difference in their relative performances but suggests the latter approach may be more robust since a mis-specification in one of the series of models will not compromise the others. This does not hold when estimating all choices simultaneously.

The important point is selecting which variables to include in the model and/or which to collect data on. Heckman et al. (1997) show that omitting important variables can seriously increase bias in the resulting estimates. A priori it is not possible to know the direction of the bias. In fact, only those variables that influence both participation and outcome should be included. It is not always clear what these variables should be, however, and careful judgement is required. Ideally, economic and social theory should provide some guidance in determining which are the important variables. An intuitive justification for only including those variables that affect participation and outcome is as follows. Variables that affect neither participation nor outcome are clearly irrelevant. However, if a variable influences only participation, there is no need to control for the differences between the treatment and the comparison group since the outcome variable of interest is unaffected. Conversely, if a variable influences only the outcome variable, there is no need to control for it since it will not be significantly different between the treatment and comparison groups. This just leaves variables that affect participation and outcome. Such variables will differ systematically between the treatment and comparison groups, significantly influencing observed outcomes in the two groups. Consequently, these are the variables that should enter the propensity score model.

However, should a variable thought to influence outcomes perfectly predict participation , problems arise. In this case, all those with a particular characteristic will either participate, or all will not participate. Clearly, with traditional matching this would result in no match being found . With propensity score matching, an analogous problem results since in this case participants will have a propensity score of 1 and non-participants a propensity score of 0. Namely, the requirement noted earlier that any combination of characteristics seen among those in the treatment group may also be observed among those in the non-treatment group is violated. The problem would manifest itself as a support problem. In these instances it may be difficult to get an unbiased estimate of programme impact using PSM.

A variation of the problem was encountered in the evaluation of the mandatory New Deal for Long-Term Unemployed people (Lissenburgh, 2001). Since the programme was piloted in specific areas, living in such an area was a perfect predictor of programme participation and if ‘area’ had been included in the propensity score model no matching would have been possible. Rather than accept this rather artificial situation, the approach followed was to exclude ‘area’ from the participation equation but to include other variables that aimed to capture area-specific characteristics. This is an intuitively appealing approach in that, while not controlling for the effect of living in a specific area, it controls for the effect of living in an area with specific characteristics. If it is these characteristics that are thought likely to influence participation and outcomes, rather than the physical location itself, including them in

the participation equation may side-step any potential bias. While this is an untestable proposition, it appears plausible.

The participation model should only include variables that are unaffected by participation, or the anticipation of participation. To do otherwise would be to mask possibly important programme effects, undermining the interpretability of estimated effects (Heckman, LaLonde and Smith, 1999). To see this, imagine that one effect of our training programme is that it raises morale (broadly defined). Should higher morale be associated with increased likelihood of success in job search, this indirect effect of the programme would be suppressed through matching were a measure of morale post-training included among the variables in the participation equation. To be certain of not obscuring important effects in this way, the variables in the participation decision should be fixed over time (e.g. gender, ethnicity) or measured before participation in the programme. In the latter case, judgement must be exercised in order to be confident that the variable has not already been influenced by the anticipation of participation.

It may be felt that, in the presence of uncertainty, it is better to include too many rather than too few variables. Again, this is a matter of judgement for the analyst. However, there are strong reasons why over-parameterised propensity score models should be avoided. First, it may be that including extraneous variables in the participation model exacerbates the support problem. Second, although the inclusion of non-significant variables will not bias estimates or make them inconsistent, it can increase their variance (Lechner and Smith, 2002). Where selection into the treatment is particularly strong, so that participants and non-participants differ markedly across observable characteristics, leaving out variables from the participation equation that play a subordinate role in the outcome equation has been suggested as a way of increasing the randomness of the selection process and reducing the variance of the matching estimates (Augurzky and Schmidt (2001).

In the spirit of this discussion it is useful to briefly mention a variant of matching that does not treat all variables equally. Puhani (1998) implements his matching estimator by finding individuals in the comparison group who are identical with respect to a number of key variables and then, within that sub-group of the comparison group, finding the individual with the closest propensity score match. This lends greater emphasis to the variables for which an exact match is required and may be appropriate when a variable is considered to be particularly important. Another motivation for this approach is that a variable may not exist in both the treatment and the comparison groups. In Puhani's case, he wishes to ensure that participants with a given length of unemployment *prior to programme entry* are matched with those in the comparison group who have the same length of unemployment.

Finally, it is worth noting that carrying out matching on sub-populations is another means of attaching additional importance to key variables. [13] To see this, consider estimating programme effects separately for men and women. This requires splitting the sample along gender lines and then performing the matching separately for either group. This is tantamount to insisting on a perfect match in terms of gender and then

[13] Heckman, Ichimura and Todd (1997) and Heckman et al. (1998) do this by implementing matching separately for four demographic groups in the national JTPA study for the US.

carrying out propensity score matching for men and women separately. However, in this scenario, two effects will be estimated – one for men, one for women. The two-stage approach followed by Puhani, on the other hand, yields a single estimate relating to the full sample.

5.2 Performing the match

The estimation of the participation model[14] results in the propensity score. This forms the basis for the match. However, before proceeding to matching, it is first necessary to ensure that any combination of characteristics seen among those in the treatment group may also be observed among those in the non-treatment group. Put differently, the common support requirement (CSR) must be enforced. The CSR must be applied differently depending on what programme effect is to be estimated. Ensuring the existence of potential matches in the non-treatment group for those in the treatment group is sufficient for the estimation of the TT. However, for ATE, it is also required that any combination of characteristics seen among those in the non-treatment group may also be observed among those in the treatment group. In other words, potential matches must exist in the treatment group for those in the non-treatment group.

Most commonly, the CSR is enforced by discarding those in the treatment group who have a probability of participation that lies outside the range evident in the non-treatment group.[15] For these discarded individuals, the programme effect cannot be estimated. Hence, unless we are prepared to assume that treatment effects are common across all individuals, the estimated effects can only be viewed as relating to that sub-population for which there is support in the non-treatment group. Where the proportion lost is small, this poses few problems. However, it has already been shown that quite substantial reductions in sample size can result. In this case, there may be concerns about whether the effect estimated on the retained individuals can be viewed to be representative of the sample as a whole. It may be instructive to inspect the characteristics of the discarded individuals since this can provide important clues to interpreting the estimated effects. Lechner (2000a) acknowledges that the standard response to the support problem is either to ignore it or to estimate programme effects within the common support. Both of these can provide misleading estimates. He develops an approach that can be used to derive bounds for the true programme effect and suggests that this should become a standard technique for assessing the robustness of the estimated programme effects.

Having enforced the CSR, the next step is to construct a comparison group from the non-treatment group. There are a number of possible ways of identifying this matched group. Most straightforward is the nearest-neighbour method. This involves taking each treated individual in turn and identifying the non-treated individual with the closest propensity score. The resulting set of non-treatment individuals constitutes the comparison group. It may be that a single non-treatment individual provides the closest match for a number of treatment individuals. In this case, the non-treatment individual will feature in the comparison group more than once. The end result is that the comparison group is of the same size as the treatment group, although the comparison group may feature fewer individuals. That is, each treated individual has

[14] For simplicity, we will assume a single treatment programme. Where issues specific to multiple treatment programmes arise they will be noted explicitly.
[15] See, for example, Smith and Todd (2000) for an alternative way of enforcing CSR.

one match but a non-treated individual may be matched to more than one treated individual. Dehejia and Wahba (1998) find that allowing for non-treatment group members to be used more than once as comparators improves the performance of the match. Furthermore, it is less demanding of the data than permitting non-treatment group individuals to be used only once. Essentially, it avoids the problem of the non-treatment group being 'used up'. Should a certain type of individual be common in the treatment group but relatively uncommon in the comparator group, the pool of comparators able to provide a close match would become exhausted were non-treatment group members used only once.

A variant on this nearest-neighbour matching is caliper matching (Cochrane & Rubin, 1973). Its defining characteristic is a tolerance it sets when comparing propensity scores, in other words a 'propensity range' in which a match is deemed acceptable. Where the propensity score of a treated individual falls beyond the bound set for a near comparator, this treated individual will remain unmatched. The appeal of caliper matching is that it imposes a form of quality control on the match. Any treatment group members left unmatched are discarded from the analysis. Hence, it represents an alternative way of imposing the CSR. However, a practical objection to its use is that it will often not be obvious how to set the tolerance.

Both nearest-neighbour and caliper matching attempt to identify a single non-treatment group member for each treatment group member. The resulting match is as good as it is possible to achieve in that the bias across the treatment and comparison groups is minimised. However, it also disregards potentially useful information by not considering any matches of slightly poorer quality. Over-reliance on a reduced number of observations can result in programme effects with larger standard errors than is strictly necessary. Nearest-neighbour and caliper matching can both be extended to allow for multiple comparators for each treatment group member (see, for respective examples, Smith and Todd, 2000 and Dehejia and Wahba, 1998).

Kernel matching also uses multiple comparators for each treatment group member. The 'kernel' is a function that weights the contribution of each comparison group member, usually so that more importance is attached to those comparators providing a better match. The most common approach is to use the normal distribution (with a mean of zero) as a kernel, where the weight attached to a particular comparator is proportional to the frequency of the distribution for the difference in scores observed. This means that exact matches get a large weight, and poor matches get a small weight. It is, however, not clear what the variance of the normal distribution (called the 'bandwidth' in this context) used should be.

In kernel matching, all members of the non-treatment group are used, to some extent, to construct a match for each treatment group member, although the contribution of those for whom the match is poor may be negligible. Other kernels would weight the relative contributions differently. The triangle kernel, for example, operates similarly to the caliper matching method in that it does not include non-treatment group members who offer an insufficiently close match. The difference from caliper matching, however, is that those who are included are weighted according to their proximity with respect to the propensity score.

Given this choice, how should one select (and justify) a particular approach? An immediate point to note is that, in theory, all approaches should yield the same results when applied to large datasets. This leaves the question of what constitutes a large dataset. However, it is reassuring that, in practice, the choice of matching method often appears to make little difference (see, for example, Smith and Todd, 2000). In small samples, the choice of matching approach can be important (Heckman at al, 1997). However, there appears to be little formal guidance in the choice of optimal method. The choice should be guided in part by what the distribution of scores in the comparison and treatment samples looks like. For example, if some treated persons have lots of close neighbours and others only have one, one would favour kernel matching or caliper matching over multiple nearest neighbour matching because the latter would result in many poor matches. Taking another example, if the comparison and treatment samples are of roughly equal size, then single nearest neighbour matching makes more sense than it does when the comparison sample is much larger than the treatment sample because in the latter case single nearest neighbour matching would result in throwing out lots of useful information. Pragmatically, it seems sensible to try a number of approaches because, as noted earlier, the performance of different matching estimators varies case-by-case and depends largely on the data structure at hand (Zhao, 2000). Should they give similar results, the choice may be unimportant. Should the results differ, further investigation may be needed in order to reveal more about the source of the disparity. This serves to reinforce the belief that matching should be implemented in a thoughtful way and not treated as a black box. More specifically, judgement and consideration is required at each stage of the process.

Finally, before moving on to an account of how to assess the performance of the match, it is worth noting that it is possible to match on a measure other than the propensity score. Lechner (2000b) and Puhani (1998) match not on the propensity score itself but on the index that underlies the propensity score.[16] The advantage of this is that the index differentiates more between observations in the extremes of the distribution of the propensity score. This means that using the index rather than the propensity score may be useful in those situations where there is some concentration of observations in the tails of the distribution.

5.3 Assessing the performance of the match

To be effective, matching should balance characteristics across the treatment and matched comparison groups. The extent to which this has been achieved can be explored in the first instance by inspecting the distribution of the propensity score in the treatment and matched comparison groups. These should appear similar. Noticeable differences should raise concern about the success of the match. However, the real purpose of matching is to balance the characteristics. With this in mind, it is very informative to inspect summary statistics for the treatment group and compare them with the matched comparison group. Slightly more formally, a measure of the bias can be calculated for each characteristic in order to achieve a standardised indicator of the degree to which the matching has been successful in balancing (see, for example, Sianesi, 2001). In the multiple treatment case, such tables quickly

[16] Technically, a probit model (for example) assumes the existence of an underlying index (or 'latent' variable). Transforming this index using the normal distribution results in the propensity score.

become unwieldy and a summary measure is more appropriate. In the case of NDYP, for example, five possible treatments means that twenty such tables would be required. Instead, Bonjour et al. (2001) consider just the average level of standardised bias in each table and present these as a summary of the performance for each pairwise match. It is probable however, that the evaluator would wish to undertake the full analysis before finalising the results.

A similar approach can be adopted to assist with specifying the participation model. The balancing test was proposed in Rosenbaum and Rubin (1983) and applied in, for example, Dehejia and Wahba (1998). The test operates by dividing the treatment group into a number of strata on the basis of the propensity score. If the characteristics of the treatment and comparison groups are not significantly different from each other within these strata then the participation model has been adequately specified to balance the characteristics. If the characteristics are not balanced for some strata, the next step is to divide the strata more finely. If the characteristics are not balanced no matter how finely the strata are defined, this indicates that it may be appropriate to introduce additional flexibility into the participation model by allowing for higher-order terms (squares, cubes etc) for particular variables or for interactions between variables. Hence, this test can provide a useful diagnostic on the specification of the participation model. However, it is important to be aware of its limitations. The balancing test does not aid in solving the variable selection problem. It only aids in model specification for a given set of conditioning variables. It cannot provide any indication as to whether the CIA is plausible, merely whether the matching has been successful in balancing the characteristics included in the model across the treatment and comparison groups. The assumption underlying matching remains a matter of faith.

Some analysts also point to the ability of the participation model to predict accurately the actual treatment status of individuals (Smith and Todd, 2000: 18). However, the ability of a propensity score model to classify correctly is only partially helpful (Lechner and Smith, 2002). For example, a model which predicts everyone to participate when only 70% do so may appear quite good on this basis when really you could argue that it has no predictive power. Indeed, it is not necessary for the propensity model to be an excellent predictor of participation. The propensity scores serve only as devices to balance the observed distribution of covariates across treated and comparison groups. Thus, 'the success of the propensity score estimation is assessed by this resultant balance rather than by the fit of the models used to create the estimated propensity score' (D'Agostino and Rubin, 2000: 750).

5.4 Considerations when using survey data

Often, programme effects will be evaluated using survey data rather than administrative records. The use of survey data is beneficial in that they are usually richer than administrative data and therefore more likely to include sufficient variables to be confident that the CIA is satisfied. However, there are also drawbacks. For reasons of survey non-response, or over-sampling of particular sub-groups, survey data will often not be representative of the population from which they are drawn. To overcome this, it is usual to apply weights to restore the profile of the sample to that of the population on a number of key characteristics. This is a standard approach with regression-based techniques. With matching, the approach is slightly more

complicated. In estimating the participation model which generates the propensity scores the analyst should use the sampling weights for both the treated and comparison group observations to restore representativeness. Post-matching, the sampling weight for each treatment group member should be applied to the corresponding comparator(s). Hence, matching using sample data results in a hybrid weight being applied to comparison group members. This hybrid weight incorporates both the matching kernel weight (if appropriate) and the sampling weight. Green et al. (2001) use sample weights in this way in the evaluation of ONE. Bryson (2001) also combines sampling weights with matching weights in an analysis of the union membership wage premium using matched employer-employee data. He finds that the application of sampling weights has a profound influence on his substantive results. Despite these empirical findings, this issue is not well covered in the literature. It appears that, even when analysts have been using sample data, they have tended to ignore weights for differential probabilities of selection into the sample and sample attrition. In these cases, one should, strictly speaking, interpret results as applying to the specific sample on which the analysis is conducted, rather than extrapolate to the population from which the sample was drawn.

Another complication when using survey data relates to the observations discarded when enforcing the CSR. Clearly, the proportion of the sample discarded may differ substantially from the proportion of the population that falls outside the common support when account is taken of sampling weights. To see this, consider sampling weights that are simply the inverse of the probability of being interviewed. This results in high weights being attached to those individuals who are under-represented in the sample. It seems plausible that those same individuals (in the treatment group) are less likely to have suitable matches in the non-treatment group and will be discarded in order to enforce the CSR. In this case, the higher weights for these individuals will mean that the proportion of the population outside the common support will be higher than the proportion of the sample outside the common support. This should be borne in mind when considering the representativeness (and, indeed, the usefulness) of the resulting estimates of treatment effects.

6. Practical considerations in using Propensity Score Matching as an evaluation tool

As noted in Section Four, all non-experimental estimation techniques make assumptions to identify programme effects and, in general, these assumptions are untestable. The violation of one or more identifying assumptions underpinning a particular technique does not necessarily mean that we should dismiss using the technique out of hand. Rather, it is important to consider the likely seriousness of the violation, the direction of any bias introduced, and the sensitivity of results to alternative econometric specifications. The theme explored in this section is the importance of understanding the assumptions being made in identifying a causal programme effect. In general, the fewer and weaker the assumptions needed to estimate a programme effect, the more realistic those estimated effects are likely to be. With these point in mind, Section 6.1 identifies circumstances in which it may be appropriate to rule out PSM as an option. Many of the situations that rule out PSM also rule out some or all of the other non-experimental estimators discussed in this paper. Section 6.2 discusses practical solutions to some of the problems raised. Section 6.3 illustrates how PSM has been used in three programme evaluations in Britain, highlighting how practical issues were identified and the solutions that were adopted. Throughout, we are pointing to considerations that researchers and commissioners of research should bear in mind when devising evaluation methodologies and interpreting evaluation results. We are not outlining rules which should be followed.

6.1 When to rule PSM out

The conditional independence assumption is untenable

Where information affecting entry to a programme and the outcome of interest is missing from the data that will be used for evaluation the key identifying assumption underpinning matching is violated, and resulting estimates of programme effects may be biased. It should be emphasised that this also applies in the case of the simple regression estimator since this shares the conditional independence assumption.

There are two general conditions in which missing information may occur:

- where insufficient time or resources have been devoted to the collection of key pre-programme variables, such as work or earnings histories;
- where it is extremely difficult to obtain information known to affect participation and outcomes, either because it is unobservable, or because the data are not available to capture it. For instance, if programme administrators are picking the best candidates for the programme ('cream-skimming') on data they observe that the analyst does not , this would seriously compromise the use of PSM, resulting in an over-estimation of programme effects. Similarly, if motivation was largely determining programme entry, the absence of pre-programme data capturing this effect could upwardly bias estimates based on PSM.

In some cases, efforts may have been made to obtain the requisite data, but data may be missing or prone to error for a sub-set of cases. One can assume that data are missing at random and simply leave out cases with missing data. However, estimates

based on a sub-sample for whom data are available may produce different results to estimates based on a broader sample. Furthermore, missing data may be non-random whereupon accounting for the patterns in missing data may be important in effecting a match between treated and non-treated groups. In these cases, propensity scores should be estimated both on the observed values of covariates and on the observed missing data indicators. D'Agostino and Rubin (2000) illustrate how this can be done.

Limited knowledge of factors affecting programme participation

Where an evaluator has little theoretical or empirical evidence on the nature of selection into a programme, it is difficult to know which factors influence participation and outcomes, and thus what set of variables to use in estimating propensity scores. Since estimates of programme effects can be sensitive to the vector of variables used, this may bias estimates. This may be a problem with innovative programmes adopting novel methods to assist clients. In this instance pre-evaluation research would be needed to identify the factors involved.

Insufficient common support

When estimating the impact of treatment on the treated, it is necessary to identify 'like' non-participants to match to participants across the propensity score distribution. If there is no common support for a substantial proportion of participants, the treatment effect is not being estimated for those individuals. With PSM, the treatment effect can only be estimated within the common support. This support may only be available for a non-representative sub-group of all participants. This presents problems from a policy perspective because, ideally, one would like an estimate of the treatment effect for all participants. The same point applies when estimating effects on the whole eligible population (the average treatment effect), but this time the common support requirement is more demanding, since it requires 'like' participants in the sample to match to non-participants.

In general, common support will be a problem when participants differ markedly from non-participants. The extent of the common support problem may not be known in advance: it may only become apparent once an expensive data collection exercise has been completed and the data analysed. However, common support problems may be apparent to those devising or administering the programme. For instance, they may be aware that only those with a very high level of motivation are volunteering for the programme. Very low take-up of a voluntary programme may be an early signal of such a problem. (Conversely, very high take-up may present difficulties if there are few non-participants from which to draw matched comparators). The fact that the evaluator possesses the information may not be enough, since the correctly estimated propensity score distributions will differ so much across participants and non-participants. In such circumstances, few observations will meet the common support requirement. In cases where it is known that one variable is a strong predictor of participation, the comparison group can be sampled on the basis of this variable.[17]

[17] This approach was proposed but not implemented in the US National JTPA Study where it was known in advance that being unemployed was a strong predictor of participation, with the implication that unemployed eligible non-participants should be over-sampled in the comparison group. Had this

In the extreme, there may be attributes that only the participants possess. If this attribute affects outcomes it will undermine PSM's ability to isolate the causal effect of treatment. This might occur if treatment was compulsory in some areas, and comparators were being drawn from non-treatment areas. In this case, treatment and area location are inseparable. Since location is a perfect predictor of treatment, it can not be used to estimate the propensity score, even though it is likely to affect outcomes. This makes the PSM estimates more vulnerable to violation of the conditional independence assumption. This was precisely the problem faced when evaluating the New Deal for the Long-term Unemployed: the problem and the way it was tackled are discussed in Section 5.1.

Small sample sizes

Small sample sizes present three problems for propensity score matching. First, they exacerbate any common support problem. Even if participants and non-participants are quite different, it is possible to estimate the impact of treatment on the treated provided the sample size of non-participants is large enough. This is because, even though there may be only a small percentage of non-participants who can be used as matches throughout the propensity score distribution, with a large enough sample, there will be some support throughout. With smaller samples, gaps appear in the common support, so that treatment effects can only be retrieved for a sub-set of the treated population. Second, small samples increase the variance of estimated effects, making identification of significant effects more difficult. Third, with fewer matches available, the evaluator may be prepared to accept more distant matches, resulting in increased bias. In these circumstances, estimated effects may be sensitive to the choice of the type of matching. This problem was faced in the ONE evaluation, and the implications and approach adopted are explained in Section 6.3.

Programme design

The primary condition that needs to be met in order for PSM to be feasible is that there should be a 'like' comparator group available to provide matches for participants. This is difficult where a programme is compulsory for an eligible group. This was the case with the New Deal for Young People.[18] Consequently that evaluation, discussed in Section 6.3, assessed the impact of parts of the programme, (the four Options and Gateway) by comparing participants in those sub-programmes with 'like' NDYP participants who had taken up other parts of the programme. In other instances, efforts have been made to assess the overall effect of a compulsory programme run in specific areas by comparing outcomes for participants with matched non-participants living outside the treatment areas. For reasons discussed above, this may be problematic.

At the other extreme, where participation is wholly voluntary, and very few of the eligible group have participated, evaluation using PSM may be problematic for reasons discussed above.

---

occurred the support problems in the Heckman et al (1998) and Heckman, Ichimura and Todd (1997) papers would have been greatly reduced. We thank Jeffrey Smith for this observation.

[18] However, for an example of an evaluation of NDYP using a non-NDYP comparison group see Blundell et al., 2001.

Where the process determining selection into a programme is not captured by the available data, PSM may produce biased impact estimates. There may be some programmes where this is more likely to occur. If individuals' participation decisions are not independent of one another, estimates of propensity scores should account for this non-independence. For example, eligible participants' decisions may be affected by early participants' programme participation. This may occur where there are limited participation slots, or where early participants' estimation of the benefits of participation feeds into others' participation decisions. Failure to account for this may bias PSM estimates of programme effectiveness.

In general, provided like comparators can be identified, there is little about the nature of the programme itself which will prevent estimation of programme effects using PSM. Indeed, one of the advantages of PSM is that it is flexible with respect to what constitutes treatment. Once a definition of participation or treatment is established, the sample can be partitioned into participants and non-participants and the matching process can begin. However, the nature of the programme and how it is configured may affect the sort of treatment effect that can be estimated. In a voluntary programme, estimates of treatment on the treated can be obtained relatively easily, but additional assumptions need to be made if the evaluator is to estimate the average treatment effect.

When random assignment is a realistic option

We discussed the advantages and disadvantages of propensity score matching relative to random assignment in Section 4. There are circumstances in which random assignment is problematic and PSM offers advantages. However, there are two key advantages of random assignment relative to propensity score matching which generally make it a superior method for obtaining unbiased estimates of treatment effects. The first is that with properly conducted random assignment there is no concern that the conditional independence assumption has not been met since we can be confident that the treated and non-treated populations are the same (within the bounds of random error) on both observable and unobservable characteristics. Second, random assignment ensures common support across the whole sample. Thus, if programme design, time, resources and ethical considerations permit random assignment, this is preferable to PSM.

When the evaluation question cannot be answered with PSM

As discussed in Section Four, there are evaluation questions that PSM cannot answer. These include those relating to the distributional effects of a programme such as the percentage of participants likely to benefit, and effects of the programme beyond the eligible population. Where general equilibrium effects have a substantial impact, partial equilibrium evaluation techniques such as PSM will produced biased estimates of programme effects. This will occur where a treatment is so substantial that it influences the general behaviour of workers or employers, such as interventions which fundamentally alter the relative price of labour across types of worker (Smith, 2000).

6.2 Designing for a PSM evaluation

Perhaps the biggest threat to accurate estimation through PSM is the time and resource pressures that evaluators face when commissioning and conducting programme evaluations. PSM is not a cheap evaluation option. It is data-hungry, requiring the collection of substantial pre-programme information to estimate propensity scores. Planning ahead is necessary if the commissioned research is to produce data which are adequate for the task of PSM. The problems are more acute for the evaluation of a voluntary programme since the matching has to be robust enough to control for self-selection bias. The two main concerns are: how to identify the factors that matter for PSM; and how best to collect data on those factors. Taking these in turn:

- For a programme that has already been, at least partially, implemented before the evaluation is to start it may be possible to carry out research to explore the reasons for participation (i.e. the factors driving self-selection onto the programme). This might be either survey or qualitative work, the latter being particularly useful if it is suspected that personal factors, such as motivation, understanding of what the programme offers, confidence, etc. are key.
- Having established the reasons for participation, and assuming these reasons cannot be captured by the data in administrative systems, data on the reasons will need to be collected on a pool of participants and non-participants from which the matched samples will be drawn. If some of the reasons for participation, such as motivation, are likely to change as a direct result of participation, and if data on these reasons cannot be collected for the pre-participation period retrospectively, then the data collection period will need to be pre-participation. This may involve data collection amongst a very large sample of the eligible population if an adequate pool of subsequent participants is to be generated. Generally speaking, the lower the participation rate, the larger the sample needed.

Although these two concerns will drive the design for a PSM evaluation there are a large number of practical issues that will inevitably arise which, as far as possible, need to be anticipated in advance. The description, in Section 6.3, of the ongoing NDLP evaluation outlines some of the issues arising there.

6.3 Examples of programme evaluations using PSM

The examples described are the evaluations of ONE, the New Deal for Young People and the New Deal for Lone Parents.

6.3.1 The evaluation of ONE

ONE is a work-focused interview for new claimants to benefits. The evaluation of its medium-term effects included an estimate of its effect on the job outcomes of participants relative to what would have happened to participants if they had not participated. In estimating this effect of the ONE treatment on the treated, participants in the four original ONE pilot areas were matched to non-participants in the same areas. Lone parent claimants and sick or disabled claimants were analysed separately. Although JSA claimants were also eligible for ONE, they were excluded because their participation was compulsory from the outset. A full account of this

evaluation can be found in Green et al. (2001). Here we shall focus on the following issues:

- Clarity about the treatment effect being estimated, and thus the interpretation to be placed upon the results
- Data availability and the credibility of the CIA in the ONE evaluation
- Matching with small samples
- Accounting for complex survey design and sample attrition when comparing outcomes post-matching.

The treatment effect estimated in the ONE evaluation

The sample was taken from the first cohort of claimants eligible for ONE, whose participants entered the programme in September or October 1999. At that time, the programme was voluntary for lone parents and sick and disabled claimants. Outcomes were measured at a second wave interview 10-11 months after participants entered the programme. In the intervening period, the programme had become compulsory for those making new claims in the sense that failure to attend an initial work-focused interview without good cause could attract benefit sanctions. To help draw causal inferences about participation in the programme, and so as to restrict participation to participation in the voluntary ONE scheme, participation was defined as those who said they had attended a work-focused interview 4-5 months after their initial claim. It is worth noting that, if those eligible for ONE were aware that it was going to become compulsory in the near future, this may have affected their labour market behaviour. For instance, those averse to participating in ONE may have been less inclined to leave benefit for a risky job if they felt that they would be subject to the compulsory scheme in the near future. Thus, the ONE impact analysed is the effect of early participation in a voluntary scheme that subsequently became compulsory for those making new claims. The effect captures actual treatment, rather than the intention to treat, since the participation definition is based on whether sample members said they have actually had an interview.

Data availability and the credibility of the conditional independence assumption

There was no intention to evaluate ONE using PSM at the time the data were collected. However, as Table 1 indicates, rich data were available from a combination of administrative and survey sources. They include many of the variables we would expect to influence participation and labour market outcomes for the two client groups, including information on labour market experiences in the two years prior to programme eligibility. However, the data contain no information on motivation prior to the programme: the only information on attitudes to working were collected after the date of eligibility and, as such, may have been affected by the ONE process. The analysis only captures motivation effects in so far as they are correlated with observed patterns of working prior to programme eligibility. This is a potential weakness of the analysis.

Table 1: Data used to generate propensity scores in the ONE evaluation

ONE-related - Pilot area - Date of eligibility	**Demographic** - Gender - date of birth - ethnicity - health
Work-experiences and qualifications - activities in 2 years pre-eligibility - social class in job pre-eligibility - highest qualification - valid driving licence	**Household** - age of youngest child - number of children - partnership - housing tenure - telephone - other household workers

Matching with small samples

There are a number of ways to identify the comparator group through PSM. Perhaps the most common is the single nearest-neighbour technique used, for instance, in the evaluation of the New Deal for Young People (Bonjour *et al*., 2001). This involves finding for each treated individual that non-treated individual with the most similar propensity score – and so, the most similar characteristics. The ONE evaluation used an alternative technique known as radius matching (Dehejia and Wahba, 1998). In radius matching all the comparators falling within the designated propensity score radius are matched to the participants in that part of the propensity distribution. Participants are matched with 'synthetic' comparators composed of a weighted equivalent of the comparators falling within the radius of their propensity score. All comparators within the radius receive equal weight.[19] Those outside the radius for all participants are not matched at all. We adopted this approach rather than nearest neighbour matching because non-participants were far more prevalent than participants. Consequently, nearest neighbour matching would have ignored those cases that were reasonable matches, although not the closest. For this reason radius matching will result in lower variances by using more data points than nearest neighbour matching, something that is particularly important when identifying a treatment effect with a small sample. The lone parent sample consisted of 379 non-participants and 171 participants: after enforcing common support, the matched data consisted of 355 non-participants and 162 participants.[20] Table 2 shows the convergence of mean values on key variables in the propensity equation, with non-participants' characteristics closely resembling those of participants after matching.

[19] However, if a particular non-participant also falls within the radius for a match with another participant, this will affect that non-participant's final weight.
[20] The restricted sample size was due to the small number of pilot areas.

Table 2: An illustration of data pre- and post-matching for lone parents in the ONE evaluation

Variable	Participants (%)	Non-participants pre-matching (%)	Non-participants matched (%)
A level or above	34	25	33
Licence	56	45	55
Any time looking after home in 2 years before eligible	47	62	51
Child < 3 years	25	36	30
2+ children	15	22	16
Benefit area:			
1	22	23	22
2	18	20	17
3	41	23	39
4	20	34	22
In 30+ hours work for 80%+ of 2 years pre-ONE	14	10	14

Results may be sensitive to the size of the radius that is the basis for matching. Choosing different radii results in non-participants being matched to different participants. Furthermore, the smaller the radius, the more difficult it is to find a match within that range, resulting in a greater number of cases failing the support requirement. We therefore tested the sensitivity of our results to three radii: 0.005, 0.01 and 0.02. In fact, results did not differ very much when using different radii.

Like any technique, results can be sensitive to the number of observations available for analysis. It is important to recognise the limitations of the matching methodology in the presence of small samples. It is possible that results may differ with a newly drawn sample. The best way to guard against this is to have large sample sizes. Unfortunately, the low take-up rate for ONE meant that these were not available for the ONE evaluation.

Accounting for complex survey design and sample attrition when comparing outcomes post-matching

The evaluation results can only be applied to the whole eligible population once sample non-response and attrition are accounted for. This is because differential non-response and attrition may lead to biases in estimated programme effects. In order to address this, sample weights were constructed using probabilistic models. The first weight attempts to correct for non-response. To do this, a probit model of survey response was estimated across all individuals in the sampling frame using administrative data used to draw the sample. The inverse of the estimated probability is used to weight back to the sampling frame. This was done separately for the lone parent group and sick or disabled clients. The second weight attempts to account for sample attrition between the first and second survey interviews, conditional on being a respondent at wave one. So a probit model of response to the second wave was estimated across all wave one survey respondents and the inverse of the estimated probability is used to weight back to the sample of wave one respondents. Again, this process was undertaken for lone parents and sick or disabled clients separately. The

final sample weight is the product of these two weights.[21] It allows the sample of wave two respondents to be regarded as representative of the cohort population.

It would have been possible to derive a single weight by directly modelling the probability in a single step of an individual in the sampling frame responding to wave two survey. However, there are two reasons why it is useful to explicitly model non-response and then sample attrition. The first is that factors affecting non-response are likely to be different from those determining attrition: this proved to be the case. Secondly, only administrative data were available to estimate non-response at wave one, whereas estimating attrition to wave two conditional on response at wave one permitted the use of rich data collected at the wave one survey.

Results

The PSM evaluation indicated that early participation in ONE had no significant effect on the employment and economic activity rates of lone parents ten months after eligibility for the programme. Better labour market outcomes, apparent in the raw data, were no longer apparent once participants were compared with 'like' non-participants. Participants' better outcomes were attributable to comparative advantages that were independent of ONE. Among sick or disabled clients, the analysis of matched data confirmed what was apparent in the raw data, namely no significant difference in the outcomes of participants and non-participants.

6.3.2 The evaluation of NDYP

The New Deal for Young People (NDYP) was introduced throughout Great Britain in April 1998 with the twin aims of helping the young unemployed into work and increasing their employability. The target group is *all* 18-24 year olds who have been claiming JSA for a period of six months or more. Other 18-24 year olds with shorter unemployment spells who are deemed to face particular obstacles to employment are also eligible. The design of the programme is as follows. First, individuals enter a period of intensive jobsearch known as the 'Gateway'. Second, they enter one of four options – subsidised employment (EMP), full-time education and training (FTET), environmental task force (ETF) or voluntary sector (VS). After this, they enter a period known as 'follow-through' which offers further support such as job-search assistance.

The evaluation of the relative effectiveness of the NDYP options is presented in Bonjour et al. (2001).[22] Dorsett (2001) used administrative data to repeat the analysis for all young men[23] entering NDYP between September and November 1998, to assess the effectiveness of the options in helping the move away from benefits. It is this latter study that is considered below since, in addition to allowing discussion of the evaluation of a multiple-treatment programme, it also examines the question of the

[21] In the case of sick or disabled clients, the final weight is the product of three weights, the third weight accounting for differential probabilities of selection into the sampling frame in the first place. This arose because low take up of the programme among the sick or disabled meant it was necessary to over-sample participants to boost their numbers in the achieved sample.
[22] See Blundell et al. (2001) for an evaluation of the Gateway.
[23] Three-quarters of NDYP participants were men.

reliability of the estimated standard errors. The complications arising from dealing with survey data (that were encountered in Bonjour et al., 2001) were similar to those already covered in the account of ONE. Here we focus on the following issues:

- Evaluating a multiple-treatment programme
- Assessing the quality of the match in a multiple treatment programme
- The reliability of estimated standard errors
- Comparing unadjusted results with post-matching results

Evaluating a multiple treatment programme

Contrary to the design of the programme, a substantial proportion of individuals remained on the Gateway for longer than the intended four months. This introduced the possibility of considering, in addition to the relative effects of the NDYP options, the effects of these options relative to prolonged Gateway treatment. The usefulness of this group is that its members can be viewed (with some qualification) as receiving no treatment at the option stage. Such evidence as is available suggests that the intensity of Gateway diminishes with time such that those on an extended Gateway can be regarded as receiving little additional attention beyond their initial Gateway experience. The approach taken in the analysis was to regard anybody staying in Gateway for more than 5 months and not proceeding onto an option as a Gateway overstayer. Hence, the Extended Gateway (EGW) was derived as a pseudo-option comprising those individuals who do not enter an option but who stay on the Gateway for this extended period

The propensity scores were estimated using a series of binary choice models. In each case, a pair of options was considered and the probability of being in a given one of those options was estimated for those who were in either of the options. In this sense, the resulting propensity scores can be seen to be conditional. Taking each option in turn, the support requirement was enforced by discarding individuals in that option whose estimated probability of being in that option lay outside the range found any of the other options. Hence, those individuals who were unsupported in the pairwise comparisons with any of the other options were discarded. This means that the treatment effects were estimated for a consistent group of individuals. The alternative of estimating propensity scores for each option using a multinomial logit or probit would have resulted in common support problems by effectively restricting the data to those individuals who might plausibly have undertaken any of the four options or remained on the extended Gateway. In comparison, the proportion of cases dropped from each pairwise comparison is low.

The effect of enforcing the support requirement is shown in the Table 3 below. The overall result was that less than 1 per cent of all observations were dropped. This level is very low and is therefore unlikely to compromise the representativeness of the results. The level of rejections was not evenly distributed across the options. Hardly any of those in FTET were discarded for reasons of support, compared with 2.7 per cent of those in EMP.

Table 3: Observations dropped due to lack of support

	EMP	FTET	VS	ETF	EGW	Total
Total	2317	6550	2483	3761	5528	20639
Dropped	62	19	43	16	46	186
% dropped	2.7	0.3	1.7	0.4	0.8	0.9

Matching was carried out for all pairwise combinations of options. This was done using the single nearest-neighbour technique with replacement: each treated individual has one match but a comparison individual may be matched to more than one treated individual. Matching with replacement in this way is less demanding in terms of the support requirement since individuals in the comparator group who would provide the closest match to a number of treated individuals remain available. Should a certain type of individual be common in the treatment group but relatively uncommon in the comparison group, the pool of comparators able to provide a close match would become exhausted were matching carried out without replacement. The drawback to matching with replacement is that the standard errors are more complicated to calculate and the variance will be higher because fewer observations are being used for the implicit comparison group.

The performance of the match

Table 4 provides some diagnostics on the performance of the match. Each cell represents the difference between members of the treatment option and those of the comparison option in the actual covariates used to model option participation. There is no available metric against which to judge the performance of the match but comparing the results to other studies suggests an adequate performance of the match.

Table 4: The performance of the match

	Comparison group:				
	EMP	FTET	VS	ETF	EGW
Treatment group:					
		Mean standardised bias among covariates			
EMP		1.9	3.0	2.2	1.9
FTET	2.2		1.8	1.9	1.6
VS	2.5	2.1		2.5	2.4
ETF	2.3	1.9	2.2		1.8
EGW	2.1	1.4	1.8	2.3	

Note: For each variable, the absolute difference in means is divided by the square root of the average of the two associated variances and multiplied by 100. Averaging across all variables yields the entry in each cell. The values in each cell can be interpreted as bias as a percentage of standard error.

Obtaining correct standard errors when interpreting results

Table 5 presents the estimated treatment effects. The outcome variable was whether the individual was claiming JSA at the time of the week commencing 26 June 2000.[24] The figures emboldened on the leading diagonal give the percentage in each option who were claiming at this time. This was lowest among those in EMP. The entries in the off-diagonal cells show, for those in the option indicated by the row, how participating in that option affected their level of claiming compared to what it would have been had they participated in the option indicated by the column. These differences are given as percentage points and a positive entry indicates that those in

[24] For the purposes of this analysis, individuals on an option were regarded as still claiming JSA.

the row option were more likely to be unemployed than they would be had they instead been in the column option.

Table 5: The estimated programme effects - levels of JSA claims in w/c 26 June 2000

q	EMP	FTET	VS	ETF	EGW
EMP	**32.06**	-19.60**	-19.78**	-19.82**	-6.83**
	-	(11.55)	(8.91)	(9.96)	(3.93)
	-	[10.87]	[8.31]	[9.93]	[3.18]
FTET	21.10**	**55.73**	-2.73	-3.02	14.24**
	(12.72)	-	(1.66)	(1.91)	(10.94)
	[12.53]	-	[1.71]	[1.95]	[9.84]
VS	19.63**	0.12	**57.38**	-3.93*	15.70**
	(9.23)	(0.08)	-	(2.03)	(9.30)
	[8.54]	[0.07]	-	[1.81]	[8.11]
ETF	16.93**	1.87	-0.08	**59.73**	15.57**
	(8.56)	(1.28)	(0.04)	-	(10.27)
	[7.74]	[1.12]	[0.04]	-	[8.90]
EGW	2.88	-16.96**	-15.20**	-13.97**	**41.34**
	(1.45)	(13.26)	(9.13)	(8.64)	-
	[1.40]	[12.34]	[8.47]	[7.87]	-

Note: approximated t-statistics appear in parentheses below the estimated effects. These were computed using the variance approximations given in Lechner (2001). The entries in brackets are the t-statistics implied by the bootstrapped standard errors (250 replications).
* significant at 5% level; ** significant at 1% level.

To illustrate, the rate of unemployment was almost 20 percentage points lower among EMP participants than among FTET participants. The results show that EMP dominates the other options. That is, members of any other option would have fared better had they instead been in the EMP, and members of EMP would have fared worse had they not participated in that option. However, the comparison with EGW shows that, while EMP participants were less likely to be unemployed than had they remained on the Gateway, the benefit to those on EGW of participating instead in EMP was not statistically significant. Overall, EGW can be seen to dominate the options other than EMP. There is little to distinguish the three other options, although those in VS would be more likely to be unemployed had they instead entered ETF. A similar result is suggested for FTET, although this fails to achieve statistical significance at the conventional level.

As noted earlier, matching with replacement results in standard errors that are difficult to calculate. The usual method of achieving reliable standard errors in such cases is to use a technique known as 'bootstrapping'. The drawback to this technique is that it is very computationally expensive. In Table 5 the effect of using bootstrapping rather than an approximation to the true standard errors can be assessed. For all the statistically significant estimated effects, the precision of the estimate drops slightly, but the two sets of t-statistics are sufficiently similar as to suggest that using approximated standard errors need not be a cause for concern in this application.

Comparing unadjusted results with post-matching results
Finally, it is worth noting that the conclusion that EMP is dominant and that EGW dominates the three other options would also have been reached through a cursory inspection of the relative mean levels of the unadjusted outcome variable. Hence, it is reasonable to ask what benefit the matching analysis has brought. Some insight into

this can be provided by considering the extent to which the estimated effects derived through matching differ from those calculated as simple differences between the average outcomes of individuals in pairs of options (i.e. estimates which do not attempt to account for non-random selection into the options). In Table 6, the entry in each cell represents the difference between the unadjusted estimate of programme effect and the estimate derived through matching. Thus, for example, the effect of being in EMP compared to FTET would be estimated simplistically at –23.7 percentage points (using the entries on the diagonal of Table 5). The matching estimate (for those in the employment option) was –19.6 percentage points. Hence, the overall effect of carrying out the matching was to reduce the apparent effect by 4.1 percentage points. In other words, once account is taken of the differences between the employment and education option participants, the estimated relative effect falls. This provides some support for the approach adopted here since some of these differences are substantial, most notably when considering ETF. In most cases, matching reduced the estimated absolute effect of treatment on the treated.

Table 6: How the matching results differ from the unadjusted differences

	EMP	FTET	VS	ETF	EGW
			Change in effect		
EMP		-4.1	-5.5	-7.8	-2.4
FTET	2.6		1.1	-1.0	0.2
VS	5.7	1.5		1.6	0.3
ETF	10.7	2.1	2.4		2.8
EGW	6.4	2.6	-0.8	-4.4	

Note: entries in grey correspond to those comparisons for which no significant effect was found at the 95 per cent level.

6.3.3 The evaluation of NDLP

The New Deal for Lone Parents (NDLP) is a voluntary programme targeted at lone parents on Income Support with children over three. Personal Advisors work with eligible lone parents with the ultimate aim of increasing the number of lone parents taking up paid work. The programme is voluntary and participation rates have been between 5 and 10%, although these have increased recently, following the introduction of compulsory regular Personal Adviser meetings for lone parents with a child aged five and over.

NDLP has been introduced nationally, but an earlier version of the programme was piloted within a small number of areas. The evaluation of that pilot generated a considerable amount of information (both quantitative and qualitative) on the factors influencing participation which the design of the national NDLP evaluation was able to draw on.

The evaluation of the national NDLP programme is perhaps unique amongst UK labour market policy evaluations in that it was specifically designed to use PSM as the main tool for estimating programme impact. A sample of those choosing to participate in the programme were to be matched to an approximately equal sized sample of non-participants with similar propensity scores. Outcomes (i.e. entries to work) are collected in face-to-face interviews, the costs of which are kept down by matching just one non-participant per participant rather than finding multiple matches.

The pilot work on the reasons for participation suggested that factors such as attitudes to childcare and to work were important influences on participation. Since these attitudinal factors can be expected to change as a result of participation in NDLP, efforts were made to collect data prior to participation. This was achieved by sending a postal questionnaire to a random sample of 65,000 eligible lone parents who, at the time the sample was selected, had not yet participated in NDLP.

The respondents to this survey, of which there were about 45,000, were monitored over time using administrative records, and subsequent participation in NDLP was recorded. After a few months the sample could be divided into three groups: those participating before completing their postal questionnaire; those participating after completing their postal questionnaire; and those not participating. The first of these three groups was discarded on the grounds that the attitudinal data captured on the questionnaire referred to the post- rather than pre-participation period. The second group formed the pool from which the participant sample was drawn (in fact, the whole of the group was selected, giving approximately 1,600 participants). The third group formed the (large) pool from which the matched sample was selected using PSM.

In order to select the matched sample, a propensity score was estimated for each member of the participant and non-participant pools using data from the postal questionnaire, benefit histories taken from administrative records, and local area characteristics (taken from the 1991 census). The propensity score model was fitted as a logistic regression model, variables being selected forward stepwise. The final model showed, as anticipated, the propensity to participate being higher than average for those:

- Actively looking for work
- Interested in gaining help or advice about work options
- Recently in contact with other local services (such as CAB, Benefits Agency local office, careers advisers etc.)
- Having a higher education qualification
- Currently studying or training
- Having worked relatively recently
- Without a limiting health condition
- Not seeing child care as an absolute obstacle to work.

Having estimated the propensity score, each of the 1,600 participants was individually matched to the non-participant with the closest propensity score using nearest neighbour matching. The very large pool (over 40,000) of non-participants from which the match could be made meant there was little, if any, support problem.

Having made the match, the participants and their matched non-participants were followed up after a period of about six months to collect outcome data. During that interview data was also being collected on factors that might have influenced participation but which could not be reliably collected within the postal questionnaire. This data will be used to ‘control’, either by re-matching the samples or using regression methods, for any residual differences in observables between the participants and non-participants. At the time of writing the data from these follow-up interviews has not been fully analysed and the results of the evaluation are not yet

available. However one early, encouraging, result is that the response rates to the follow-up interview are almost identical in both the participant and matched comparison groups.

Although the evaluation of NDLP probably represents the most comprehensive effort to date in the UK to make PSM a reliable and accurate method for estimating a programme impact, a number of potential difficulties, fully acknowledged by the evaluators, still exist which could bias the estimates:

1. Although the postal questionnaire collected data on all the reasons for participation that the evaluators were aware of, there is still no guarantee that *all* the reasons were covered by the questionnaire or that the questions were answered reliably enough by respondents. Thus, even after matching on the propensity score estimated using all of the rich data available, the CIA may still not have been met.
2. There is inevitably a time lag between completing the postal questionnaire and participation. But for each individual the evaluators must assume that the reasons for participation are captured by the postal questionnaire. In reality, for some individuals, the decision to participate will be based on a change in circumstances since completing the questionnaire. For these individuals the evaluators will be unaware of the change in circumstances and the propensity score estimated will be inaccurate. This leads to violations of the CIA, the seriousness of which cannot readily be established (although the face-to-face follow-up interview does include questions that attempt to establish for participants whether a change in circumstances triggered their participation).

7. Summary and Conclusions

This review of evaluation techniques and the role of matching has focussed on the assumptions that evaluators have to make to identify the causal effect of a programme on labour market outcomes where that programme (or 'treatment') is not randomly assigned. To know the effect of programme participation on employment or earnings, we must compare the observed outcome for participants with the outcome that would have resulted had that person not participated. However, this counterfactual outcome cannot be observed and this is the reason for the evaluation problem. Seen in this way, the essential difficulty in estimating programme effects is one of missing data. Causal inference is problematic because, where people who participate differ systematically from those who do not participate in ways which might affect their labour market prospects independently of participation, we can not infer the outcomes from non-participation for participants simply by comparing participants' outcomes with those of non-participants.

We have described how statistical matching using the propensity to participate can be used to construct the counterfactual. The technique continues to attract attention as a useful evaluation tool in the absence of random assignment. The method has an intuitive appeal arising from the way it mimics random assignment through the construction of a control group post hoc. Results are readily understood, but less so the assumptions underpinning the validity of the approach and the circumstances in which those assumptions are violated. There are no 'magic bullets' in non-experimental programme evaluation: rather, there are a range of techniques designed to estimate programme effects which are based on differing identification assumptions. We have been at pains to outline the assumptions underpinning PSM and conditions in which they seem most plausible, and where they seem particularly implausible. The key assumption made in PSM is that selection into a programme can be captured with observable data that are available to the evaluator. This is known at the Conditional Independence Assumption (CIA). For this identifying assumption to be plausible, one must be able to control for all characteristics affecting both participation and the outcome of interest. This requires very informative data on individuals prior to programme eligibility. This presents a challenge to commissioners of evaluation research since the painstaking collection of such data is often at odds with the tight scheduling of evaluation research. However, without these data, PSM is likely to produce biased estimates.

The CIA is an untestable assumption, so that one can never be sure whether one has correctly specified the analysis to capture variables influencing participation and outcomes. However, all non-experimental estimation involves making generally untestable assumptions to identify causal effects. The likely violation of one or more of these assumptions does not mean that we should dismiss using the technique out of hand. But it is important to be clear about the likely seriousness of the violation, the direction of any bias introduced, and the sensitivity of results to alternative specifications. At a minimum, this places the onus on the evaluator to test the sensitivity of results to different matching approaches, and then compare them to other techniques to test for implicit selection biases.

References

Agodini, R. and Dynarski, M. (2001) *Are Experiments the Only Option? A Look at Dropout Prevention Programs*, Mathematica Policy Research Inc., Princeton NJ

Augurzky, B. and Schmidt, C. M. (2001) *The Propensity Score: A Means to An End*, IZA Discussion Paper No. 271, Bonn

Bergemann, A., Fitzenberger, B. and Speckesser, S. (2001) 'Evaluating the Employment Effects of Public Sector Sponsored Training in East Germany: Conditional Difference-in-Differences and Ashenfelter's Dip', mimeo, University of Mannheim

Black, D., Smith, J. Berger, M. and Noel, B. (2000) *Is the threat of reemployment services more effective than the services themselves: experimental evidence from the UI system* Unpublished manuscript, University of Ontario.

Blundell, R. and Costa Dias, M. (2000) *Evaluation methods for non-experimental data* Fiscal Studies 21(4).

Blundell, R., Costa Dias, M., Meghir, C. and van Reenen, J. (2001) *Evaluating the employment impact of a mandatory job search assistance programme* IFS working paper wp01/20.

Blundell, R. and MaCurdy, T. (1999) *Labour supply: a review of alternative approaches* in Ashenfelter, O. and Card, D. *Handbook of labor economics vol III*, Amsterdam.

Bonjour, D., Dorsett, R., Knight, G., Lissenburgh, S., Mukherjee, A., Payne, J., Range, M., Urwin, P., White, M. (2001) *New Deal for Young People: national survey of participants: stage 2* Employment Service Report ESR67.

Bryson, A. (2001) *The Union Membership Wage Premium: An Analysis Using Propensity Score Matching*, Centre for Economic Performance Working Paper No. 1169, London School of Economics

Burtless, G. (1995) 'The Case for Randomized Field Trials in Economic and Policy Research', *Journal of Economic Perspectives*, 9(2), 63-84

Cochrane, W. and Rubin, D. (1973) *Controlling bias in observational studies: A* Review, *Sankhya, Series A*, 35: 417-46

D'Agostino R.B. Jr. and Rubin, D.B. (2000) 'Estimating and using propensity scores with partially missing data'. *Journal of the American Statistical Association*, 95(451): 749-759.

Dehejia, R. and Wahba, S. (1998) *Propensity score matching methods for non-experimental causal studies* NBER working paper 6829.

Dehijia, R. and Wahba, S. (1999) 'Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs', *Journal of American Statistical Association*, 94(448), 1053-1062

Dorsett, R. (2001) *The New Deal for Young People: relative effectiveness of the options in reducing male unemployment* PSI Discussion Paper Number 7.

Eichler, M. and Lechner, M. (2000) 'Some Econometric Evidence on the Effectiveness of Active Labour Market Programmes in East Germany', William Davidson Institute Working Paper No. 318

Friedlander, D. and Robins, P. K. (1995) 'Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods', *The American Economic Review*, Vol. 85, No. 4, 923-937

Friedlander, D., Greenberg, D. H. and Robins, P. K. (1997) 'Evaluating Government Training Programs for the Economically Disadvantaged', *Journal of Economic Literature*, Vol. XXXV, December, 1809-1855

Frölich, M., Heshmati, A. and Lechner, M. (2000) *A microeconometric evaluation of rehabilitation of long-term sickness in Sweden* University of St. Gallen Department of Economics discussion paper 2000-04.

Gerfin, M. and Lechner, M. (2000) *Microeconometric evaluation of the active labour market policy in Switzerland* Discussion paper, University of St. Gallen

Goldberger, A. (1983) *Abnormal selection bias* in Karlin, S., Amemiya, T. and Goodman, L. (eds.) *Studies in econometrics, time series and multivariate statistics* New York: Academic Press.

Green, H. Connolly, H., Marsh A., and Bryson, A. (2001) *The Longer-term Effects of Voluntary Participation in ONE*, Department for Work and Pensions, Research Report Number 149

Heckman, J. (1995) *Instrumental Variables: A Cautionary Tale*, NBER Technical Working Paper No. 185, Cambridge, Mass.

Heckman, J. and Hotz, .V. J. (1989) 'Choosing Among Alternative Methods of Estimating the Impact of Social Programs: The Case of Manpower Training', *Journal of the American Statistical Association*, 84(408), 862-874

Heckman, J., Ichimura, H. and Todd, P. (1997) 'Matching as an econometric evaluation estimator: evidence from evaluating a job training programme' *Review of Economic Studies* 64: 605-654.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998) 'Characterizing selection bias using experimental data' *Econometrica* 66(5): 1017-1098.

Heckman, J. and Robb, R. (1985) *Alternative methods for evaluating the impact of interventions* in Heckman, J. and Slinger, B (eds.) *Longitudinal analysis of labor market data* New York: Cambridge University Press

Heckman, J., Lalonde, R. and Smith, J. (1999) *The economics and econometrics of active labour market programs* in Ashenfelter, O. and Card, D. *Handbook of labor economics vol III,* Amsterdam.

Heckman, J. and Robb, R. (1986) *Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes* in Wainer, H. (ed.) *Drawing inferences from self-selected samples* Berlin: Springer-Verlag.

Heckman, J. and Smith, J. (1995) 'Assessing the Case for Social Experiments' *Journal of Economic Perspectives* 9 (2) 85-100

Heckman, J. and Smith, J. (1999) 'The pre-programme earnings dip and the determinants of participation in a social programme. Implications for simple programme evaluation strategies' *Economic Journal* 109: 313-348.

Heckman, J., Smith, J. and Clements, N. (1997) 'Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts', *Review of Economic Studies*, 64(4), 487-537

Imbens, G. (1999) *The role of the propensity score in estimating dose-response functions* NBER technical working paper 237.

Imbens, G. and Angrist, J. (1994) 'Identification and estimation of local average treatment effects' *Econometrica* 62(4).

LaLonde, R. (1986) 'Evaluating the Econometric Evaluations of Training Programs with Experimental Data', *American Economic Review*, 76,(4), 604-620

Larsson, L. (2000) *Evaluation of Swedish youth labour market programmes* Office of Labour Market Policy Evaluation (IFAU) working paper 2000:1.

Lechner, M. (2000a) *A note on the common support problem in applied evaluation studies* University of St. Gallen Department of Economics discussion paper 2000-01.

Lechner, M. (2000b) 'An evaluation of public-sector-sponsored continuous vocational training programmes in East Germany' *Journal of Human Resources* 35: 347-375.

Lechner, M. (2001a) *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption* in Lechner, M. and Pfeiffer, F. (eds) *Econometric evaluation of labour market policies* Heidelberg: Physica-Verlag.

Lechner, M. (2001b) 'Programme heterogeneity and propensity score matching: an application to the evaluation of active labour market policies' *Review of Economics and Statistics* (forthcoming).

Lechner, M. and Smith, J. (2002) 'Some Exogenous Information Should Not Be Used in Evaluations', mimeo

Lissenburgh, S. (2001) *New Deal for the Long Term Unemployed Pilots: quantitative evaluation using stage 2 survey* Employment Service Research Report ESR81.

Puhani, P. (1998) *Advantage through training? A microeconometric evaluation of the employment effects of active labour market programmes in Poland* ZEW discussion paper 98-25.

Puhani, P. (2000) 'The Heckman Correction for Sample Selection and Its Critique – A Short Survey', *Journal of Economic Surveys* 14, 53-68

Purdon, S. (2002) *Estimating the impact of labour market programmes*, Department for Work and Pensions Working Paper No. 3
Rosenbaum, P. and Rubin, D (1983) *The central role of the propensity score in observational studies for causal effects* Biometrica 70: 41-50.

Rubin, D. B. (1974) 'Estimating Causal Effects of Treatments in Randomised and Non-randomised Studies, *Journal of Educational Psychology*, 66, 688-701

Sianesi, B (2001) *An evaluation of the active labour market programmes in Sweden* Office of Labour Market Policy Evaluation (IFAU) working paper 2000:5.

Smith, J. (2000) 'A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies', Schweiz. Zeitschrift fur Volkswirtschaft und Statistik, Vol. 136(3) 1-22

Smith, J. and Todd, P. (2000) *Does matching overcome LaLonde's critique of nonexperimental estimators?* mimeo, downloadable from http://www.bsos.umd.edu/econ/jsmith/Papers.html.

Vella, F. (1998) *Estimating models with selection bias: a survey* Journal of Human Resources 33(1).

White, M. and Kileen, J. (2000) *The Impact of Careers Guidance for Employed Adults in Continuing Education*, PSI Research Discussion Paper No. 5.

White, M. and Lakey, J. (1992) *The Restart Effect: Evaluation of a Labour Market Programme for Unemployed People*, London: Policy Studies Institute