

# **ESTIMATING THE IMPACT OF LABOUR MARKET PROGRAMMES**

**Working Paper No 3**

**Susan Purdon**

# **ESTIMATING THE IMPACT OF LABOUR MARKET PROGRAMMES**

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

**By**

**Susan Purdon  
National Centre for Social Research**

© Crown copyright 2002. Published with permission of the Department Work and Pensions on behalf of the Controller of Her Majesty's Stationary Office.

The text in this report (excluding the Royal Arms and Departmental logos) may be reproduced free of charge in any format or medium provided that it is reproduced accurately and not used in a misleading context. The material must be acknowledged as Crown copyright and the title of the report specified. The DWP would appreciate receiving copies of any publication that includes material taken from this report.

Any queries relating to the content of this report and copies of publications that include material from this report should be sent to: Paul Noakes, Social Research Branch, Room 4-26 Adelphi, 1-11 John Adam Street, London WC2N 6HT

For information about Crown copyright you should visit the Her Majesty's Stationery Office (HMSO) website at: [www.hmsogov.uk](http://www.hmsogov.uk)

First Published 2002

ISBN 185197 966 2

ISSN 1476 3583

## Contents

<b>1</b>	<b>THE EVALUATION PROBLEM</b> .....	<b>1</b>
<b>2</b>	<b>WHY ESTIMATE THE COUNTERFACTUAL?</b> .....	<b>3</b>
<b>3</b>	<b>DEFINING THE COUNTERFACTUAL</b> .....	<b>5</b>
<b>4</b>	<b>ESTIMATING ADDITIONALITY - THE COMPARISON GROUP PROBLEM</b> .....	<b>7</b>
<b>5</b>	<b>AN OVERVIEW OF THE MAIN DESIGNS</b> .....	<b>9</b>
<b>6</b>	<b>WHO AND WHAT IS TO BE MEASURED</b> .....	<b>13</b>
<b>6.1</b>	<b>Defining the ‘programme group’ for the evaluation</b> .....	<b>13</b>
<b>6.2</b>	<b>Defining the outcomes</b> .....	<b>15</b>
<b>7</b>	<b>RANDOMISED TRIALS</b> .....	<b>19</b>
<b>7.1</b>	<b>How randomisation solves the evaluation problem</b> .....	<b>19</b>
<b>7.2</b>	<b>Implementation</b> .....	<b>19</b>
7.2.1	The eligible population for the trial .....	20
7.2.2	Sample sizes and sub-groups .....	21
7.2.3	Gaining informed consent.....	22
7.2.4	The randomisation procedure.....	22
7.2.5	The ‘alternative’ programme.....	23
7.2.6	Collecting outcome data on the programme and control groups.....	24
<b>7.3</b>	<b>What randomised trials cannot reliably answer</b> .....	<b>25</b>
<b>7.4</b>	<b>Why randomisation is not always used</b> .....	<b>26</b>
<b>7.5</b>	<b>Randomisation of areas</b> .....	<b>26</b>
<b>8</b>	<b>QUASI-EXPERIMENTAL DESIGNS</b> .....	<b>29</b>
<b>8.1</b>	<b>Before-after designs</b> .....	<b>29</b>
8.1.1	Design issues .....	30
8.1.2	Problems in the interpretation of before-after designs.....	32
8.1.3	Before-after designs with voluntary programmes .....	34
<b>8.2</b>	<b>Interrupted time-series designs</b> .....	<b>34</b>
8.2.1	The analysis of time series data .....	35
8.2.2	Problems.....	35

<b>8.3</b>	<b>Difference-in-differences enhancements .....</b>	<b>36</b>
<b>8.4</b>	<b>Other time-series approaches.....</b>	<b>37</b>
<b>8.5</b>	<b>One-to-one matched comparison group design .....</b>	<b>38</b>
8.5.1	Collecting data on the factors that influence participation .....	40
8.5.2	Selecting the programme group.....	41
8.5.3	Selecting the comparison group/propensity score matching.....	41
8.5.4	Problems in the interpretation of matched comparison designs .....	42
8.5.5	Collecting data on the outcomes from the programme and comparison groups .....	43
8.5.6	When matched comparison designs are appropriate .....	43
8.5.7	Matched comparison group designs used in combination with other quasi-experimental methods.....	43
<b>8.6</b>	<b>Statistical modelling of existing data to evaluate voluntary programmes .....</b>	<b>44</b>
8.6.1	Propensity score matching with kernel weighting.....	44
8.6.2	The instrumental variables estimator .....	45
8.6.3	The Heckman selection estimator.....	45
<b>8.7</b>	<b>Matched area comparison design.....</b>	<b>45</b>
8.7.1	Selecting the pilot areas.....	46
8.7.2	Selecting the comparison areas.....	46
8.7.3	Problems in the interpretation of matched area comparison designs.....	47
8.7.4	A difference-in-differences enhancement.....	48
8.7.5	A possible improvement for voluntary programmes .....	48
<b>9</b>	<b>SAMPLE SIZE CALCULATIONS .....</b>	<b>51</b>
<b>9.1</b>	<b>Sample size calculations for randomised trials.....</b>	<b>51</b>
<b>9.2</b>	<b>Sample size calculations for randomised area trials .....</b>	<b>54</b>
<b>9.3</b>	<b>Sample size calculations for the main quasi-experimental designs.....</b>	<b>55</b>
<b>10</b>	<b>EXAMPLES OF THE MAIN EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS .....</b>	<b>57</b>
<b>11</b>	<b>BIBLIOGRAPHY .....</b>	<b>61</b>

## **ACKNOWLEDGEMENTS**

I would like to thank Vivienne Avery and other members of Analytical Services Division within the DWP for their advice, assistance and support in writing this paper.

## **THE AUTHOR**

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 a feasibility study for a randomised trial of job retention and rehabilitation services.

# 1 THE EVALUATION PROBLEM

Labour market programme evaluations in Britain are typically large, complex and expensive, but a considerable proportion of the time, money and effort is devoted to estimating just one figure, namely the ‘impact’ of the programme. This paper describes the main designs available for making this estimate, the assumptions needed for each design, and the circumstances under which each design might be appropriate. An overview of more general evaluation methods, including process evaluation, is given in Purdon et. al. (2001) Research Methods for Policy Evaluation. DWP Working Paper No. 2.

The ‘impact’ of a programme is usually phrased in terms of ‘additionality’, that is, the number of additional positive ‘outcomes’ that the new programme produces. Additionality is itself measured as the difference between two numbers:

the outcomes that occur under the new programme

*minus*

the outcomes that would have occurred without the programme.

The first of these two numbers is relatively easy to estimate. Typically it will be estimated using administrative or survey data.

The second figure, which is known as the *counterfactual*, is generally very difficult to estimate. This paper is devoted to methods of estimating this figure.

The reason for the difficulty of measurement is simply because individuals cannot be in two states at once – if individuals are exposed to a programme then we cannot observe what would have happened to them if they had not been exposed. So estimating the counterfactual involves making an estimate about a hypothetical state – what would have happened if the programme had not been introduced.

In general, the counterfactual is not of interest in its own right. The main focus of interest is usually the estimate of additionality. However, since estimating additionality once the counterfactual is known is a relatively trivial extra step, evaluation designers tend to concentrate on the counterfactual rather than additionality per se.



## 2 WHY ESTIMATE THE COUNTERFACTUAL?

The primary reason why so much effort is put into estimating the counterfactual and additionality is that it gives the best direct evidence of whether or not a programme 'works'. For instance, if a programme is designed to help the unemployed find work, then additionality will be defined in terms of the extra numbers finding work under the new programme relative to whatever help was available previously. If the estimate of additionality is positive (so that more people find work with the programme than without) the programme will be judged to 'work'. If in addition, additionality is greater than some threshold figure than the programme might be judged not only to 'work' but also to be cost-effective.

It should be noted that measuring additionality is not the only interest of evaluators. Although an estimate of additionality *will* give some measure of whether a programme 'works' the estimate does not help greatly to explain why the programme works, and in many instances, unless separate estimates of additionality are available, it will not be possible to say who the programme works for. To answer these sorts of questions, evaluations to estimate additionality ('impact evaluations') are usually carried out alongside process evaluations which look in detail at how programmes operate.



### 3 DEFINING THE COUNTERFACTUAL

As was noted above, the counterfactual can be defined as:

‘the outcomes that would have occurred if the new programme had not been introduced’.

This definition, although simple enough to state, includes, however, a number of ambiguities that need to be thought through in deciding on an evaluation strategy.

*(1) What is the alternative state that the new programme is to be compared against?*

In most instances additionality will be relative to any current programmes that are targeted at the same eligible population. In this case the counterfactual could be rephrased as

‘the outcomes that would have occurred under the old programme’.

Sometimes additionality is to be measured relative to a state where no alternative programmes exist. In this case the counterfactual could be rephrased as

‘the outcomes that would have occurred without a programme’.

In other, much rarer, instances there are two (or more) new programmes (say, A and B) to be tested against each other. In this case the ‘counterfactual’ could be rephrased as

‘the outcomes that occur under programme B’.

More commonly, where there are two or more new programmes to be tested each of these will be compared to the currently existing programme, and the counterfactual is then written in one of the first two ways described above.

In many labour market evaluations the definition of the counterfactual state will need very careful consideration. For example, if a very intensive version of a New Deal programme was to be introduced into a few areas, the counterfactual could be defined in terms of a ‘pre New Deal’ state, or in terms of the standard New Deal programme. Assuming the standard New Deal *does* have an impact on outcomes, additionality under the former definition will be higher than in the latter. Since, as will become clear in this paper, large additionality figures are easier to detect than small figures, the counterfactual written in terms of ‘pre New Deal’ is probably easier from an evaluator’s standpoint. But if the question of interest to policymakers is whether the intensive New Deal is better than the standard New Deal, then the counterfactual needs to be written in terms of the standard New Deal, and the evaluators have to work with that.

*What are the ‘outcomes’?*

The counterfactual is defined in terms of ‘outcomes’. These are discussed separately in Section 6.2.

*(2) What population does the counterfactual refer to?*

Most programmes have a target or eligible population and the additionality of a programme will usually be written in terms of this group. There are however, particular problems for labour market programmes because of stock and flow issues (where the stock is the eligible population of existing cases at the start of a new programme and the flow are the newly eligible cases that arise after the start of the programme). In addition there are often other sub-groups for which separate estimates of additionality are either needed or are desirable. (Usually these will be sub-groups of the eligible population but, in principle, there could be other, non-eligible, groups which the programme is expected to have an impact on.)

For voluntary programmes some thought needs to be given as to whether the population is all the eligible population or the volunteering population.

These issues are considered further in Section 6.1.

*(4) What point in time does the counterfactual refer to?*

Defining the counterfactual as

‘ the outcomes that would have occurred if the new programme had not been introduced’

suggests that a decision is needed on the time interval between the start of a new programme and the time when outcomes are to be measured. Alternative timings for the outcome measures can lead to quite different estimates of additionality.

This is discussed in Section 6.2.

## 4 ESTIMATING ADDITIONALITY - THE COMPARISON GROUP PROBLEM

Having defined the counterfactual, the additionality attributable to a new programme is, in very simplistic terms, written as:

$$I = Y_1 - Y_0$$

where  $I$  = impact or 'additionality' ;

$Y_1$  = 'outcomes' under new programme

$Y_0$  = 'outcomes' under the counterfactual.

As was noted earlier, the most difficult part of this equation is the estimation of the counterfactual (i.e.  $Y_0$ ). Given that the counterfactual is purely hypothetical, the way evaluators have tackled this problem is to 'construct' a sample or population who will experience the counterfactual programme. This sample is usually called the 'control' group or 'comparison' group. Most controversies about evaluations stem from disagreements about the suitability or otherwise of the chosen comparison group.

Although there are exceptions, the general procedure for making the estimate of additionality is as follows:

1. Decide which groups of the population estimates of additionality are needed for.
2. Define the outcomes that are of interest.
3. Identify a group who will experience the counterfactual (i.e. the 'control' or 'comparison' group).
4. Apply the new programme to all or a sub-sample of the eligible population (typically this group will be called the 'treatment' or 'programme' group).
5. Measure outcomes on the programme group (i.e. calculate  $Y_1$ )
6. Measure outcomes on the comparison group (i.e. calculate  $Y_0$ ).
7. Calculate additionality as the difference between  $Y_1$  and  $Y_0$ .

As noted above, the most difficult, and most controversial step of this procedure is Step 3, the choosing of an appropriate comparison group.

The reason why the choosing of a comparison group is difficult is because of the fact that the outcomes of interest are always related to factors beyond being exposed to the new programme. For instance, if a new programme is designed to help people into work, the outcomes of interest will be written in terms of movements into work. For an individual any such movement will be influenced by a whole range of factors including qualifications, previous work history, local economic circumstances, and motivation. Over and above these factors the new programme may, or may not, be an influential factor.

The comparison group will only give a convincing estimate of the counterfactual if this group has the same mix of people, in terms of personal characteristics such as qualifications, motivation etc., and in terms of local economic conditions, as the programme group. In other words the comparison group should look identical to the programme group on all the factors that influence outcomes, with the sole exception that the programme group is exposed to the new programme and the comparison group is not. Comparison groups are controversial when there is some suspicion that there are systematic differences between the programme and control groups, and that these differences bias the estimate of additionality.

## 5 AN OVERVIEW OF THE MAIN DESIGNS

The main designs for estimating the counterfactual fall into two main groups: experimental methods, which are essentially randomised trials, and quasi-experimental methods. The quasi-experimental methods themselves fall into two main sub-groups: designs that can be used on programmes that are introduced nationally at one point in time, and designs that can only be used on programmes that are to be piloted before full implementation. In addition there are quasi-experimental methods that are appropriate for voluntary programmes. All of the methods that are suitable for the evaluation of national programmes can also be applied to pilot programmes.

The main designs are described very briefly below. Each of the designs is discussed in more detail in Section 8.

### *Randomised trials*

The randomised trial is the gold standard against which all quasi-experimental methods are judged. The main feature of a trial is that the eligible population is divided, completely at random, into two groups: a programme group and a control group. The power of the randomised trial lies in the fact that the randomisation ensures that the two groups are balanced in terms of all factors that can affect outcomes, with the single exception that the programme group are exposed to the new programme and the control group are exposed to the 'alternative' programme. No other evaluation design can guarantee this balance between the programme and comparison group.

Randomised trials are not, however, without problems and there are a number of ways in which biases might occur. Some of the most common problems are described in Section 7.

The comparison group for a randomised trial is usually referred to as the 'control' group. For all non-randomised designs it is conventional to use the term comparison group rather than control group.

### *Before-after designs*

Before-after designs are most commonly used to evaluate programmes that are to be introduced nationally at one point in time, although they can be used to evaluate pilot programmes as well. The comparison group in this instance is drawn from the eligible population before the programme is implemented, whereas the programme group is drawn from the eligible population post-programme implementation.

The main disadvantage of the before-after design is that change, or additionality, due to the programme can not be disentangled from change that might naturally occur between any two points in time. Because the two groups are selected from

different periods of time the groups cannot be guaranteed balanced on factors that affect outcomes (over and above the programme itself), especially for factors, such as labour market conditions, that are known to vary over time. The design gives results that are very difficult to interpret if the programme is introduced around the same time as other related programmes.

#### *Interrupted time-series designs*

An improvement on the basic before-after design is to increase the number of 'before' estimates and, if possible, the number of 'after' measurements to give a time-series. An 'interruption' in the time-series is then checked for at the time, or shortly after, the programme is implemented. The interruption is interpreted as the impact of the programme.

This design is more robust than the basic before-after design to 'natural' change over time, especially if the programme is expected to make a large impact (relative to the noise in the series). The design is not however, robust when the programme of interest is introduced around the same time as other related programmes. The design can also be difficult to use if the impact of the design occurs gradually rather than immediately.

#### *Difference-in-differences enhancements*

Both the basic before-after design, and the interrupted time series design, can be made more convincing by comparing change over time for the group targeted by the new programme with change over time for a similar group who are not eligible for the programme. This latter group can provide very useful information on 'natural' change over time against which the programme eligible population can be compared. The design is most convincing if the two populations can be shown to have always moved in parallel in the past, but that with the introduction of the programme they have moved apart. The main reason the approach is not used all that frequently is that finding a suitable group against which to compare the programme eligible population is rarely easy.

A variation on this approach is to compare change over two different outcomes rather than different groups, where the first outcome is expected to be influenced by the programme and the second isn't, yet where there is some evidence that the outcomes react similarly to other outside influences. Again, the main reason this is rarely used in practice is because of the difficulty of finding suitable outcomes against which to compare the main outcome variable.

#### *One-to-one matched comparison group design*

For voluntary programmes with low participation rates the time series approaches discussed so far often fail because the comparison groups all come from 'before' periods at a point when nobody actively volunteered for the programme (for the

simple reason that there was no programme to volunteer for). This means that time-series approaches almost always focus on additionality for the whole of the eligible population rather than on additionality amongst participants. But, because additionality for the whole population will typically be very small, the programme impact can be very hard to detect. The one-to-one matched comparison group design takes a different approach.

In the one-to-one matched comparison group design both the programme and comparison groups are selected from the post-programme implementation population. The programme group is selected from those who participate in the programme; the comparison group is selected from those who choose not to participate. The participants selected for the evaluation are matched, one-to-one, to non-participants with similar characteristics. The intention is that the matching process ensures that the two groups are very similar in terms of factors that affect outcomes – in other words the matched non-participants are selected to mimic the control group from a randomised trial.

The main objection to the method is that the matching process is rarely, if ever, as good as it needs to be, and there will always be residual differences between the participants and their matched non-participants, beyond simply being a participant. These residual differences can lead to severe bias in the estimate of additionality.

#### *Statistical modelling of existing data*

The one-to-one matched comparison group design is appropriate when collecting data on outcomes is expensive. This might happen if outcome data had to be collected by survey or by expensive to collect administrative data. If outcome data on non-participants can be collected cheaply, or at no expense, as would be the case if outcome data was available from standard administrative systems, creating a small comparison group from a large pool of non-participants would be a waste of data. In such instances it makes more sense to treat the whole of the non-participant pool as, in some sense, the comparison group, and then control for differences between participants and non-participants using statistical methods.

A number of methods have been proposed in the econometrics literature. These include propensity score matching with kernel weighting, the Heckman selection model, and the method of instrumental variables. All of these methods need very strong assumptions for unbiased estimation of the counterfactual, and the methods tend to incur the same interpretation problems as the standard one-to-one matched comparison group design.

#### *Matched area comparison design*

For programmes that are to be piloted within a limited number of areas before national implementation the most common evaluation design is the matched area

comparison. Under this design the areas that the programme is piloted in are matched to non-pilot areas with similar characteristics. The eligible population is then monitored over time within the pilot areas and their matches, and any differences in outcomes is attributed to the programme.

A variation on this basic design is where the pilot areas are compared to the rest of GB instead of to individual matched areas. In this instance outcomes for the eligible population within the pilot areas are compared to outcomes for the rest of the national eligible population (usually after modelling for differences in area-level factors such as local unemployment rates).

This design works well if the impact of the programme is expected to be large, in which case any observed difference between the programme and comparison groups will be largely attributable to the programme and not to small differences between areas. The design is problematic for programmes with only a small expected impact (which is generally the case for voluntary programmes with fairly low participation rates) since in these cases the programme impact can be swamped by the 'natural' differences between areas. Consequently, matched area comparisons work best with mandatory programmes.

Matched area comparisons are very natural candidates for a difference-in-differences approach. If time series data on outcomes can be collected in both the pilot and control areas, and if these two time series can be shown to have moved in parallel for at least a period before the programme was introduced into the pilot areas, taking a before-after difference in both the pilot and comparison areas and then comparing these differences can give a far better indicator of the programme effect than the straightforward difference between the areas. This approach is especially useful if the matching of areas is only approximate (since then there may be significant 'before' differences between areas).

## 6 WHO AND WHAT IS TO BE MEASURED

As was noted earlier, although the most difficult part of evaluation design is the definition of the comparison group, three necessary precursors are the decisions about who the programme group for the evaluation are, how the outcomes of interest are to be defined, and how data on those outcomes are to be measured. None of these are as straightforward as might at first appear, and how they are defined can quite strongly influence the subsequent evaluation design. The programme group and the outcomes are taken in turn in this section.

### 6.1 Defining the 'programme group' for the evaluation

How the programme group for an evaluation is defined depends in large part on who measures of additionality are required for. Issues to be taken into account include:

- Should all of the eligible population be represented in the programme group?
- Are there sub-groups of the eligible population for whom separate estimates of additionality are required?
- Who exactly are the eligible population?

Taking these in reverse order:

*Who exactly are the eligible population?*

For many mandatory programmes the eligible population is easy to identify – it is simply all those targeted by a programme. However, for mandatory programmes where it takes time to work through the whole of the eligible population it might sensibly be assumed that the programme cannot have any real impact on those cases that have not yet been 'processed'. So for some mandatory programmes the eligible population *for the evaluation* might be those dealt with within a fixed time period.

For voluntary programmes the issues about how to define the eligible population are more complex. It might be argued that in many instances that a voluntary programme can only change outcomes for those who volunteer to take part (i.e. there are no substitution effects). If this is the case it would be legitimate in a voluntary programme to define the eligible population *for evaluation purposes* as those who volunteer for a programme. However, it would also be legitimate to define the eligible population as all those targeted by the programme. Or it would be possible to take a position part way and identify the eligible population as all those expressing an interest in participation, or all those who had been invited to participate.

These decisions are not neutral because sample size considerations depend very crucially on them. Sample sizes are dealt with more thoroughly in Section 9, but as

an example, suppose that a voluntary programme had a participation rate of 10%. Then additionality of, say 5 percentage points amongst participants would translate into additionality of just 0.5 percentage points amongst the whole of the eligible population. An evaluation that focussed on the participants would be far more likely to detect statistically significant additionality than would an evaluation that focussed on the general eligible population. Nevertheless the decision is still not straightforward - evaluation designs that concentrate on participants rather than the eligible population have particular implementation problems. Furthermore, many of the quasi-experimental evaluation designs are ruled out if a decision is made to only look at participants.

*Are there sub-groups of the eligible population for whom separate estimates of additionality are required?*

To understand exactly who a new programme works for it is useful to be able to make separate estimates of additionality for sub-groups of the eligible population. This can help to focus the programme at a later date. For instance, in New Deal for Lone Parents (NDLP) it makes sense to ask the question whether personal advisors have more impact on lone parents with older children than lone parents with very young children. If this is found to be the case then a decision might be taken after the evaluation findings are reported to focus the attention of personal advisors onto parents with older children.

The difficulty for evaluators is in identifying the relevant sub-groups in advance of the evaluation, primarily because, if separate estimates for sub-groups are needed, the 'programme group' has to be selected in such a way that sufficient numbers of the relevant sub-groups are included.

In most instances a compromise will have to be reached between the desire to increase the numbers per sub-group, to allow for several sub-groups to be analysed separately, and the need to keep the evaluation to a manageable size.

One key categorisation that arises in most labour market evaluations is the categorisation between stock and flow, the stock being those members of the eligible population who have been eligible for some time, and the flow being those cases relatively new to the eligible population. It is almost always desirable to make separate estimates of additionality for the stock and the flow, firstly because the two estimates are likely to be very different (the stock often being more resistant to new programmes than the flow), but secondly because, over time, the profile of cases that the new programme will deal with will change: assuming cases only have a finite life on a programme, after a period of operation the proportion of cases that are 'flow' will increase. If inference about the future impact of a new programme is to be made it is vital that separate predictions of additionality for flow cases can be made.

*Should all of the eligible population be represented in the programme group?*

Estimates of additionality will not be available for any groups of the eligible population excluded from the 'programme group'. So, for instance, if the programme group is defined as those volunteering for a programme, no estimate of additionality will be available for those who don't volunteer. If additionality for this group can safely be assumed to be zero then this omission would be acceptable.

There may, in addition, be other groups who for practical or theoretical reasons are left out of the programme group. These might include:

- Groups for whom it is suspected additionality will be very small;
- Groups for whom it is difficult or impossible to collect outcome data on (this would include non-respondents to surveys);
- In a randomised trial setting, those groups who refuse to give, or cannot give, informed consent (such as, possibly, those with learning difficulties, or those who are very risk averse).
- Those entering the eligible population after the evaluation has started. (Typically, for practical reasons, the population eligible for an evaluation will be defined as those eligible within a fixed period.)

In the design of an evaluation careful consideration needs to be given to the effects of omitting such groups, especially if there is any expectation that additionality will be different for any of these groups.

## **6.2 Defining the outcomes**

Additionality is defined in terms of the difference in outcomes between the programme and comparison groups, and it follows that the estimate of additionality quoted will depend crucially on the outcome measures adopted.

Broadly speaking the outcomes for an programme will be defined in terms of (a) what is measured, and (b) when they are measured. An example will best illustrate why this is the case.

Suppose a new programme is designed to increase the number of unemployed people on benefits entering work. Then clearly the evaluation outcome measure for an individual has to include a measure of whether or not s/he has entered work. But it would be legitimate for the additionality question to be posed in any of the following ways:

- (i) at fixed time T1 is the percentage of the programme group in work greater than the percentage of the comparison group in work?;
- (ii) at fixed time T2 (later than T1) is the percentage of the programme group in work greater than the percentage of the comparison group in work?;

- (iii) over the time interval T3 has the average time spent in unemployment been shorter for the programme group than the comparison group?.

And so on. Posing the additionality question in any of these ways could potentially lead to very different conclusions about the effect of a programme.

For example, suppose the objective of Programme 1 is to reduce the duration of unemployment amongst a group most of whom will get jobs anyway if left alone for long enough. Under this scenario, the best outcome measure would probably be (iii): an outcome measure along the lines of (i) would only identify the short term effects of the programme (which would tend either to exaggerate the overall programme effect, or to detect no effect at all if the programme was slow to take off), whereas an outcome variable along the lines of (ii) could emphasise only the long term effects of the programme (in which case it might well be concluded that the programme had no effect at all).

If one of the aims of the programme was to get people not only into work, but into sustainable work, the best initial route for many people may be a spell of training or another measure to improve employability. If this isn't recognised, and the outcome for the evaluation is written as (i), the programme might actually appear to be detrimental. In this instance, (ii) would be a better way to express the outcomes, although it is possible that even better expressions could be thought of that would properly describe the complexity of the process.

Generally speaking, however the outcomes for the programme are defined, it is sensible to collect data on any changes in labour market status for a reasonably long period between the start of the programme and a finish date. This allows for various hypotheses about the effect of the programme to be tested. The finish date should be chosen at a point when it is believed the long-term effects of the programme can be identified. In practice, the need to get evaluation results out quickly often means that only short-term programme effects are estimated.

#### *Sources of outcome data*

How the outcomes are defined can have very serious implications for the costs of an evaluation. If outcomes are defined in such a way that they can be derived from administrative records then that is clearly a much cheaper option than the alternative of carrying out a survey. In practice, this means that evaluations with outcomes derived from administrative records can use much larger sample sizes than evaluations based on survey data. This in turn means that evaluations with outcomes derived from administrative data give more precise estimates of additionality.

In practice the outcomes of interest for labour market programme evaluations are entries into work or training, neither of which are fully known from administrative sources. In these instances there is a balance to be struck between the advantages of being able to report additionality in terms of the outcomes of real interest, but

having to put a fairly large margin of error around this estimate because of the limited sample size, and being able to report on additionality far more precisely but in terms of outcomes that are not of primary interest (e.g. additionality might be written in terms of exits from benefit rather than entries to work, or using exits to work drawn from incomplete administrative records which are potentially less reliable, and which omit data of interest such as occupation and earnings).

As well as the extra precision given by administrative data there is also the advantage that the data are usually not subject to the non-response biases that surveys suffer from.

If surveys *are* needed to collect outcome data then some consideration should be given to who to include in the survey. For example, suppose we have a programme and comparison group, and data is to be collected by survey on entries to work. Then one option would be to include all of the members of the two groups in the survey. An alternative option would be to only include the members of either group who have left benefits, on the grounds that most entries to work will be amongst this sub-sample. This *could* reduce the size of the survey quite considerably<sup>1</sup>, but at the possible expense of reducing detection of any ‘additional’ moves into work that don’t involve a move off benefits. However, given that with a fixed survey budget, this method could allow for quite a considerable increase in the sample size of the programme and comparison groups, this option is worth serious consideration. However, this method does reduce the usefulness of the survey data for other analyses. For instance, the survey would no longer include a representative sample of programme participants and so data on their participation experience could not be reported on without a special boost sample.

#### *Sample size estimation*

We have argued above that the outcome data for labour market programme evaluations needs to be fairly complex, with preferably full work-benefit histories being collected for a reasonable period after the start of a programme. At the planning stage however, it makes sense to think of the outcome data in more simplistic terms so that sample sizes can be calculated relatively easily. The simplest approach is to reduce the outcome variables to a simple binary variable for the sample size calculations (e.g. whether at a particular point in time a group member is or is not in work). This might be done at two or more hypothetical points in time, so that the short, medium and long term effects of the programme are considered.

#### *Reporting*

Reporting on programme effects is by no means straightforward. Many programmes will have complex effects, with, say, short term effects being different from medium term effects, which are in turn different to the long term effects. It

---

<sup>1</sup> For example, if 50% of the programme group left benefits, and of these 80% went into work, giving an overall percentage entering work of 40%, then a sample size of 1000 of those leaving work would give a standard error around the 40% of 0.63%. To achieve a standard error of this size from a survey representative of the *whole* of the programme group would need a sample size of 6000.

will not, in most instances, be possible to condense all of this complexity into a single measure of additionality. Generally a compromise will be needed between clarity and accuracy, although the exact figures that get reported may in part be driven by the needs of the cost-benefit analysis. A non-technical summary of the evaluation will typically present the results in terms of additional positive outcomes at one or more points in time (e.g. 'at one year into the programme the programme had helped 3000 people into work who would not otherwise have found work'). For outcomes based on continuous variables, such as wage levels, it may be preferable for reporting to change to a binary variable (e.g. 'at one year into the programme the programme increased the wages of participants by an average of £50 per week' or 'at one year into the programme the programme increased the wages of 20% of participants by more than £30 per week').

It should be noted that analyses of evaluation data that involve statistical modelling of the data will not automatically produce estimates of additionality as described above. To derive these estimates the usual procedure is to use the models to *predict* outcomes for the programme group in the absence of the programme (i.e. predict the counterfactual for the programme group). The difference between the predicted and observed outcomes is then the measure of additionality.

## 7 RANDOMISED TRIALS

### 7.1 How randomisation solves the evaluation problem

As has been noted earlier, the largest problem faced by evaluators is the identification of a suitable comparison group from which the counterfactual will be estimated. The main criterion that a suitable comparison group has to meet is that it is the same as the programme group on all factors that relate to outcomes, with the single exception that the programme group are exposed to the programme and the comparison group are not. So if the outcome variable is entry into work, and the programme group are young unemployed people aged 18-24, then the comparison group should be young unemployed people aged 18-24. And if the programme group includes many motivated people with reasonably high qualifications, then the comparison group should include the same proportion of motivated people with reasonably high qualifications. And so on. If there are any differences between the programme and comparison groups then there will be a strong possibility that the difference in outcomes observed between the programme and comparison groups will be partly due to this difference, and not entirely due to the programme.

This very stringent requirement for the comparison group can only be met in two ways: either information is gathered about all the factors that might influence outcomes (such as age, motivation, qualifications, labour market conditions etc.) and the comparison group is then 'constructed' so as to ensure balance between the programme and control groups on all of these factors. This method underlies all of the quasi-experimental methods. Alternatively, the comparison and programme groups can be artificially constructed from the eligible population, by *randomly* dividing this population into two groups. This random assignment to groups guarantees balance (within the bounds of random error) between the two groups. For example, if 20% of the eligible population are highly motivated, then, after randomisation, (close to) 20% of those assigned to the programme group will be highly motivated, and (close to) 20% of those assigned to the comparison group will be highly motivated.

Thus random assignment automatically gets around the problem implicit in all quasi-experimental methods, of having to know, and control for, all factors that affect outcomes except for the potential influence of the allocation itself. For this reason randomised trials are thought of as the gold standard against which other evaluation methods are judged.

### 7.2 Implementation

To implement a randomised trial the following steps will need to be followed:

1. Decide on the eligible population for the trial.

2. Calculate the sample size needed for the trial, and identify a subset of the eligible population of this size. This sub-set may need to over-represent sub-groups if separate estimates of additionality are needed for these groups.
3. If necessary collect informed consent.
4. Randomly allocate the subset to two groups: the programme group and the comparison group (known in a randomised trial setting as the control group). The allocation may be in the ratio 50:50 but does not have to be.
5. Expose the 'programme group' to the programme. Expose the control group to the 'alternative'.
6. Follow up both groups over time to collect data on outcomes. The difference in outcomes between the two groups is the estimate of additionality.

Each of these steps is discussed in turn below.

### ***7.2.1 The eligible population for the trial***

For mandatory programmes the population eligible for the trial will usually be straightforward to identify – it will simply be those eligible for the programme. If however, there are groups within this population for whom an estimate of additionality is not needed these could be excluded. This might include those who could not possibly participate fully during the lifetime of the evaluation.

In some instances the individuals in the eligible population will be identifiable at the start of the trial from administrative records. In others the eligible population will need to be defined in more abstract terms. For instance, the eligible population might be those who start to claim unemployment benefits within the next six months. In this case the eligible population is well defined but the members of the population cannot be pinpointed in advance. This has implications, that are discussed below, for how the randomisation is physically carried out.

Defining the eligible population for voluntary programmes is more problematic. There are at least two possible populations: all those eligible for the programme, and all those who volunteer. There are also several intermediate groups, such as those actually targeted, or those who express an interest in volunteering (some of whom may subsequently choose not to). If it is reasonable to assume that additionality will be zero, or very close to zero, for those who do not volunteer for a programme, sample size considerations suggest that the ideal 'population' for the trial will be those who volunteer. This however has considerable practical implications, with volunteers who are randomised to the control group having to be turned away from a programme. Nevertheless, if the volunteering rate is low, randomisation of the volunteers is probably the only practical way forward, the sample sizes needed for a randomised trial of the whole of the eligible population usually being far too large to be feasible. The only exception might be if outcome data could be collected from administrative records so that large sample sizes do not impact greatly on the evaluation costs. (See Section 9 for an example of the sample size calculations.)

## **7.2.2 Sample sizes and sub-groups**

The sample size calculations for a randomised trial are described in Section 9.

It should be noted that the calculations will need to take into account the expected percentage who will give informed consent to the random allocation if this is needed (see Section 7.2.3) and the percentage from whom it is possible to collect outcome data. This will be much less than 100% if the outcome data is to be collected by survey.

For trials where the individual members of the trial population can be identified at the design stage the actual sample size can be fixed. For trials where the eligible population is identified over time (e.g. those who in the next six months start to claim unemployment benefit, or those who in the next six months volunteer for a programme) some estimates will be needed of the likely 'flow' into the sample. The sampling fraction to be applied, or the length of time the trial will have to run for, to accrue the necessary sample size will then be determined by these estimates.

If separate estimates of additionality are needed for sub-groups (e.g. the stock and flow) the trial may need to ensure a large enough sample size for these groups. This can be achieved in two ways. Most easily the total size of the trial can be increased to the point where the smallest sub-group has a sufficient sample size. This however can be somewhat wasteful. The alternative method is to adjust the allocation ratio for different sub-groups.

To see how this works note that in many labour market evaluations it will be desirable to allocate as few people as possible to the control group. This means that for many trials the number randomly allocated to the programme will be much greater than the number randomly allocated to the control group. However, since, from a statistical viewpoint there is little advantage in having an imbalance between the two groups, if outcome data is expensive to collect, the programme group can be sub-sampled from at the outcome stage to give equal numbers in both the programme and control groups.

Suppose as an illustration that a population divides into two sub-groups, A and B, A covering 2,000 persons and B covering 10,000. Further suppose that the sample size calculations for the trial suggest that for both groups the programme group and control group need to have 1,000 persons (so that the total size of the trial is 4,000). Then one way to achieve this would be to randomly allocate Group A in the ratio 50:50, giving the 1,000 per group as required. For Group B, however, the allocation to programme and control group would be in the ratio 90:10. This would give 9,000 in the programme group and the required 1,000 in the control group. However, to save money, only a random sample of 1,000 of the 9,000 would be followed up for outcome data. So, although the trial includes 12,000 persons, only 4,000 would be included in the outcome data collection stage and subsequently in the analysis.

It should be noted that, although this method of using different allocation ratios is sound from a theoretical standpoint, it can be hard to implement in some practical situations. For instance, if informed consent is needed it is easier for staff to explain a trial where all potential trial members have a fixed chance of being in the control group, rather than to have to vary their message depending upon the sub-group the person belongs to. The method of varying allocation ratios may work better in a setting where informed consent is not needed.

A final note. Although over-representation of sub-groups within a randomised trial is usually done in order to allow for separate estimates of additionality to be made, there is also an argument for over-representing sub-groups where additionality is expected to be higher than average but where the counterfactual is relatively low. This can make the overall estimate of additionality more precise. (NB A similar argument is often used when deciding on the target population for a new programme. To increase the efficiency of a programme there is a strong argument for concentrating it on those who are most likely to benefit from it.)

### **7.2.3 *Gaining informed consent***

For voluntary programmes where randomisation occurs at the point of participation it will be necessary to get informed consent from potential participants before they enter the trial. In most instances this will involve providing potential participants with a written explanation of what the trial is about and what the implications of taking part (or not taking part) will be. This will usually be followed by a verbal explanation. Those participants who consent to enter the trial will sign a form giving that consent.

Unfortunately the procedure of getting informed consent can, in some instances, bias a trial, simply because it can have the effect of excluding certain sub-groups of the population. For instance the procedure is likely to deter those who are risk averse, and those who have difficulty understanding the issues. If the programme would have a different impact on the excluded groups than the included groups then the trial estimates will be biased. At a minimum, data on the reasons for refusing consent should be collected. This information can then be used by others to judge the validity of the trial results.

### **7.2.4 *The randomisation procedure***

The procedures for randomisation can be fairly straightforward if the trial population can all be identified at the start of the trial. For large trials the randomisation can be very simple, randomisation being done using a random number generator on a computer or, as is often done, based on a person's last NINO digit<sup>2</sup>. For smaller trials random differences between the programme and

---

<sup>2</sup> In some labour market pilot programmes this approach has been introduced for purposes other than the evaluation with, for instance, the stock being asked to join the programme in NINO order. This means that for the first few months of the programme at least, the programme is implemented as a randomised trial, although this 'natural trial' can be seriously contaminated if self-referral to the programme is allowed.

control groups can be reduced by randomising within strata. The ideal strata will be based on variables that are known to be related to outcomes. For instance, in a programme aimed at moving the unemployed into work, stratification by length of unemployment spell would be appropriate, since this is correlated with the likelihood of finding work.

For trials where the trial population is identified over time (either centrally or locally), more complex systems may need to be put into place for the randomisation. The simplest method would be to base the randomisation on a predetermined number, such as NINO, and in most instances this may be sufficient.

One difficulty with the NINO approach however is that it is theoretically open to abuse by those implementing the randomisation. Suppose for example, the randomisation is to be carried out by those administering the programme. Then it will be advantageous to the administrators to ensure 'promising' cases get into the programme group and 'unpromising' cases get into the control group. With a NINO approach the administrator would know the allocation before informed consent was obtained. At that point if a 'promising' case was destined for the control group the administrator could steer the case towards refusing consent; and vice versa for an unpromising case. If administrators are able to manipulate the procedures in this way, however subtly, the results of the trial will be biased.

To avoid this happening randomisation should only happen after consent is gained. One possibility would be for an administrator to contact a central point after consent is given and the randomisation would then be carried out at that central point. If some information is gained about the person then at this point it may be possible to incorporate an element of stratified randomisation. For the reasons noted above, the rules being used for the randomisation must not be apparent to the administrator.

### **7.2.5 The 'alternative' programme**

Careful consideration needs to be given to what 'alternative' programme will be offered to the control group since this will define the counterfactual. Ideally, in most labour market programme evaluations, the control group would continue much as they would have if the new programme had not existed. There are a number of reasons why this may not happen however:

- (1) the control group may hear about the new programme and this may affect their behaviour (they may, for instance, delay their job search activities until they become eligible for the new programme);
- (2) the control group may be offered the 'best' currently available services as an alternative to the programme. If without the new programme no procedures exist for informing people about these services the control group will not be comparable with the current situation;

- (3) the programme may have effects that indirectly affect the control group, for instance by changing the attitudes of local employers.

Related problems are 'implementation bias' and 'queuing bias'. Implementation bias will occur if, in a trial setting, a programme cannot be implemented in the way it would be in a more natural setting. For instance, in a trial of a voluntary programme the advertising of the programme may be different to the advertising of the programme under non-trial conditions. If this difference changes the profile of participants the trial results will, arguably, not apply to the full implementation case.

Queuing bias can also be a problem. Because in a trial only a proportion of the population are offered the new programme, this can give an unfair advantage to those in the programme group, relative to the control group, that would not happen under full implementation of the programme. For example, in a trial of a programme where people are helped into work, those allocated to the programme would have an unfair advantage in the situation where the number of available jobs was finite. On the other hand, if the new programme is an alternative to an existing service with limited availability the degree to which the control group can access the existing service will be increased.

All of these potential biases are very difficult, if not impossible, to quantify. Evaluators need to make efforts to minimise the biases wherever possible, but they should also ideally be able to make informed statements about how problematic the biases are likely to be.

### **7.2.6 Collecting outcome data on the programme and control groups**

As has been noted at various points, outcome data tends to be either collected through administrative sources or by survey (typically using face-to-face interviewing). In either case missing data on outcomes can cause problems of bias or interpretation. Taking each type of data in turn:

#### *Administrative data*

To ensure comparability between the programme and control groups it is important that the administrative data on both groups comes from the same source. Assuming that can be guaranteed, the only additional problem will probably be the quality of the administrative data.

Whilst errors in outcome data are not biasing as long as the errors are equally likely to occur in both groups, care must be taken in the presentation of results. As an illustration, suppose in a trial where the main outcome variable is finding work, suppose 1000 find work in the control group and 2000 in the programme group (so that the estimate of additionality is 1000 extra people in work for those in the trial). But suppose this outcome is missing in 10% of cases. Then additionality will be estimated as just 900 extra people in work. In this case the data quality is the same

on both sides (and the ‘relative risk’ of entering work equals 2 with or without the error, so the errors are not strictly speaking biasing), but the direct estimate of additionality is affected. This suggests that in instances such as these the focus should be on the estimation of *proportional* additionality rather than on the raw numbers. This, however, would have implications for cost-benefit analyses.

### *Survey data*

The problems with using error-prone administrative data for outcome data do not disappear if surveys are used instead, although in some instances (certainly not all) the quality of data collected by survey will be higher than that recorded on administrative systems. Surveys however, are problematic because they can suffer from differential non-response (which administrative records tend to suffer much less from, although it can still arise).

There is a particular problem if the response rate is higher within, say, the programme group than the control group, and if the response rate is related to outcomes. Suppose for instance, those finding work have a lower survey response rate than those not finding work. Then the additional positive outcomes in the programme group may be underestimated if a proportion of those with positive outcomes fail to respond.

One way the problems of survey non-response might be partially dealt with is in non-response adjustment at the analysis stage (which might be done by applying non-response weights). To do this data needs to be collected on the characteristics of those who don’t respond. This might be done in a trial setting by carrying out a baseline survey at the start of the trial to collect background information on those entering the trial. This would however be an expensive option, and might paradoxically make matters worse by creating additional non-response at the outcome stage. A preferable source of data might be administrative records (which can act as a good source of calibration data) or data on the trial members collected by the trial administrators at the time of recruitment.

### **7.3 What randomised trials cannot reliably answer**

Even with all the sources of potential bias listed in the previous paragraphs, randomised trials are still usually considered the most reliable method of estimating programme additionality as long as reassurance can be given that the biases will not be large. Trials do not, however, allow for all evaluation questions of interest to be answered. Their main problem is that they give an estimate of *net* additionality but they give no means of disentangling this into gross change components, that is, the numbers for whom the programme improves outcomes and the numbers for whom the programme worsens outcomes. Nor, amongst all those in the programme group, can the additional positive outcomes be identified separately to the dead-weight cases (i.e. the positive outcomes that would have

happened anyway). This severely limits the extent to which the particular benefits of the programme on individuals can be identified.

In addition the trial cannot be used to estimate additionality on those who experience intermediate outcomes. For example, it might be of interest to know whether those in a trial who enter work stay in work longer. For this analysis it would be desirable to be able to compare those who enter work in the programme group with those who enter work in the control group. But, since at this stage, the two groups are no longer comparable (the programme having had an impact on the profile of those from the programme group who enter work) this estimate cannot be reliably made. The question would have to be re-phrased as 'are those in the programme group more likely to enter longer term posts than those in the control group?' which can then be asked of the whole of the programme and control groups. In practice evaluation researchers do ask the former question but to answer it they revert to quasi-experimental methods.

It should also be noted that randomised trials of individuals are of little benefit if an objective is to bring about a general 'culture change' across the whole of an eligible population. In these instances it would be almost impossible to avoid contamination of the control group. For programmes of this type randomisation of areas (see Section 7.5) may be the only feasible way to conduct a randomised trial.

#### **7.4 Why randomisation is not always used**

Although there has been a change in recent years, traditionally labour market programmes in Britain have not been evaluated using randomised trials. There are several reasons why this is so.

- (1) There are disagreements about whether randomising people to a labour market programme is ethical.
- (2) Even if ministers and policy makers can be convinced of the benefits, it can be very hard to explain to the press and public.
- (3) The management and administration of a trial is difficult and expensive. (Two programmes have to be run within single areas at the same time. Procedures have to be put into place to gain informed consent and to randomise people. Monitoring systems have to be put into place to check that the randomisation procedures are adhered to.)
- (4) There are very valid concerns that the potential biases in trials will outweigh the benefits.

And perhaps most importantly,

- (5) Until relatively recently it was considered by many that the quasi-experimental methods described in the rest of this paper would give reasonably reliable results.

#### **7.5 Randomisation of areas**

In all the discussion of randomised trials so far it has been assumed that it will be individuals that are randomised rather than some other unit. But many of the

objections to randomised trials identified in the last section disappear if randomisation is done at an area level rather than at an individual level. The way this would work is as follows:

1. Calculate how many areas are needed for the trial. (This would be based on knowledge about the size of areas, in terms of the eligible population, and the degree to which areas differ from one another.)
2. Select the relevant number of areas. These would preferably be chosen at random, but it is not an essential requirement.
3. Divide the areas into pairs with similar characteristics. This would be based on characteristics such as the eligible population historically having had the same profile of outcomes. Where data on this is not available, pairing on general labour market characteristics would be adequate (as long as these characteristics are thought to correlate with the outcomes of interest for the eligible population).
4. Randomise one area within a pair to the programme group and the other to the control group.
5. Within each area follow up all or a sub-sample of the eligible population to collect outcome data (either by survey or through administrative records).

Although randomisation of areas, rather than individuals within areas, is very powerful and avoids many of the problems of randomised trials, it has rarely, if ever been used in labour market programme evaluations for (at least) two reasons:

- (1) The sample size calculations will inevitably suggest that the number of areas for the trial needs to be quite large. (The reason for the large sample size is because the difference between programme and control groups now incorporates random between-area differences as well as random individual differences. To ensure that 'additionality' can be detected over and above the random differences needs large sample sizes. This is akin to the clustering effects in surveys, although the problem is more acute because the cluster sizes are large.) Whilst the large number of areas is problematic in itself, a related issue is that, to get a large enough number of areas will often mean using small area geographies. But administering programmes within small areas can lead to implementation problems, such as the need to employ specialist staff within areas that are too small to keep those staff fully occupied.
- (2) For voluntary programmes it is not possible to randomise at the point of participation, since nobody participates in the control areas. This usually means that to detect additionality requires extremely large sample sizes of individuals within the randomised areas. This is not likely to be a problem if outcome data is to be collected from administrative records, but is a problem for outcome data collected by survey.

In practice a quasi-experimental method (matched area comparison) that is related to a 'randomisation of areas' method has been used instead. Although not implicit to the design, this has tended to be carried out with a fairly small number of areas

(far fewer than would be required for a randomised trial). The matched area comparison design is discussed in Section 8.7.

## 8 QUASI-EXPERIMENTAL DESIGNS

If a randomised trial approach is either ruled out or is considered undesirable the challenge for evaluators is to choose an alternative quasi-experimental design that will give results that are reasonably robust.

As was noted earlier in this paper, the designs divide into two main sub-groups: designs appropriate for programmes that are introduced nationally at one point in time; and designs appropriate for pilot programmes. In addition there are quasi-experimental methods that are appropriate for voluntary programmes.

The designs considered in this paper are:

- before-after design
- interrupted time series design
- difference-in-difference enhancements
- one-to-one matched comparison group design
- statistical modelling of existing data for the evaluation of voluntary programmes
- matched area comparison design.

The occasions when each of these designs might be useful is shown in the table below. Boxes with a tick indicate whether a design might be considered, boxes with a question mark indicate instances where under special circumstances the design might be considered but where generally the design would be inappropriate. However, in each case the appropriateness of the approach would have to be considered very carefully.

	National programmes		Pilot programmes	
	Mandatory	Voluntary	Mandatory	Voluntary
Before-after	✓	?	✓	?
Interrupted time-series	✓	?	✓	?
Difference-in-differences	✓	?	✓	?
Matched comparison	X	✓	✓	✓
Statistical modelling	X	✓	✓	✓
Matched area comparison	X	x	✓	?

### 8.1 Before-after designs

In a basic before-after design, the eligible population for a new programme is identified both before and after the programme is introduced. A 'programme group' is selected from the eligible population after the programme is introduced, and a 'comparison group' is selected from the eligible population before the programme is introduced. Outcomes are then collected on both groups and the

difference in outcomes between the groups gives the estimate of additionality. It is crucial for the design that eligibility in the before period can be established.

### **8.1.1 Design issues**

A key design issue for before-after studies is whether a cohort of the eligible population should be followed across the before and after periods, in which case the programme and comparison groups are the same people, or whether two fresh cross-sectional samples should be drawn, one before and one after. In instances where either design is appropriate, the cohort design will give more precise estimates of additionality, because random differences between the profiles of the two groups will be ruled out.

The decision between the two approaches (cohort or two cross-sections) depends primarily on how eligibility for the programme is defined. If the eligible population is static over time then a cohort design is appropriate; if the eligible population is non-static then cross-sectional samples will be appropriate. For example, to evaluate a programme targeted at the general population, a cohort design would be appropriate because, within acceptable limits, the general population before a programme is implemented will be the same as the general population after the programme is implemented. However, to evaluate a programme targeted at the unemployed, the two cross-sectional approach would be needed because the population of unemployed before a programme is implemented will not be the same as the population of unemployed after implementation

In some instances, where populations change but only slowly, there may be an argument for taking a cohort approach as long as the before and after data is collected over a short interval of time. For example, a programme offering assistance to the self-employed might be evaluated by selecting a cohort of the self-employed just before the implementation of the programme and collecting outcomes, and then returning to the cohort after the implementation of the programme to collect outcome data a second time. This would probably be an acceptable design because, over short periods of time, the population of the self-employed is *reasonably* static.

Some care needs to be taken for programmes where outcomes are affected by ageing. For example, to evaluate a health programme aimed at reducing the number of hospital stays of the elderly, one approach would be select a cohort of the elderly some months before the programme is introduced and then monitor hospital stays over the next few months. Then, after the programme is introduced, the same cohort would be monitored for hospital stays for a further period. Clearly, comparing the before-after outcomes for individuals within the cohort would not be appropriate in this instance because change over time would be observed simply because the individuals have aged by at least six months, and hospital stays would, on average, increase irrespective of whether or not the programme was introduced. In this instance the cohort design is still appropriate,

but rather than calculating change for *individuals* before and after, the analysis would compare members of the cohort who were the same age before and after. For example, a person aged 72 in the before period would be compared with a person aged 72 in the after period. This would generally be done by controlling for age in the analysis rather than by physically matching individuals, but the principle is the same.

Another issue that needs to be considered is the timing of the before and after measurements, both in terms of when the cohort and/or cross-sectional samples will be selected, and when outcomes will be measured. As a general rule the gap between the before and after outcomes should be as short as possible so that there is less opportunity for other events to occur in the interim period that might obscure the programme effect. But the exact timing depends on a number of other factors:

- Some outcome measures can be collected at the time the sample is selected (income levels would be an example). Other outcome measures need an interval after sample selection before they can be observed (an example would be moving out of unemployment into work). In the first case the 'before' sample could be selected just before the programme is implemented, but in the second case the sample has to be selected early enough to allow a suitable time for 'before-programme' outcomes to be measured.
- The timing of the 'after' outcome measures needs to be long enough after a programme has been implemented for the programme to have had an effect. There may be an argument for collecting outcome data several times in the 'after' period so that the short, medium, and long-term impacts of the programme can be estimated, but if this is to be done the 'before' period *may* have to be equally long and have the same outcome measurements taken so that a before-after comparison can be made each time. (This would not be necessary when the eligible population stays the same over time – in this instance all of the short, medium, and long-term outcomes can be compared to a single 'before' measure of outcomes, the comparison being made on a comparable group each time.)

A final point. One great advantage of cohort designs in instances where outcomes are to be collected by survey is that, if *retrospective* data on outcomes can be reliably collected, then there is no necessity to speak to the cohort members before the programme is implemented. For instance, if a programme to increase the incomes of the self-employed is introduced, a sample of the self-employed might be selected after the programme is implemented. If that sample can reliably report on their incomes both now and before the introduction of the programme then an estimate of the impact of the programme can be derived. In practice, the survey outcome data that can be reliably collected retrospectively is fairly limited so the design may not prove useful very often. An exception might arise if survey data can be linked to administrative data such as tax returns.

### **8.1.2 Problems in the interpretation of before-after designs**

The main objection to before-after designs is that change in outcomes between the before and after periods might, in principle at least, be attributable to three things:

1. Change because of the introduction of the programme
2. Change because of differences in the profiles of the before (comparison) and after (programme) groups;
3. Parallel historical change (either 'natural' change or change because of the introduction of other new policies at about the same time).

The first of these is the change that we are trying to measure; the other two elements of change are 'noise' that needs to be eliminated.

Taking the two extra components of change in turn:

#### *Change in group profile*

Suppose, as an illustration, that a programme is targeted at getting the unemployed into work, and to test the impact of the programme a sample of the unemployed is selected six months before the programme is introduced and entries to work over the next six months are monitored. Then, after the programme is introduced, a second sample of the unemployed is taken, and entries to work over the next six months are monitored again.

In this example, for the before sample to be a reliable comparison sample for the after (programme) sample, the two samples should not differ on factors that are related to entry to work. So, at a minimum, the two samples should have the same age and sex profile, and they should have the same proportion of long term unemployed.

This can usually be dealt with at the analysis stage, either by controlling for these factors in a regression analyses, by weighting the samples to a common profile, or by matching the individuals in the 'after' sample with individuals in the 'before' sample who have similar characteristics. This was attempted with some success in the evaluation of the introduction of Jobseeker's Allowance (JSA) where the 'before-JSA' sample was adjusted using weighting to resemble the 'after-JSA' sample in terms of local labour market conditions using data from the Labour Force Survey.

It should be noted however that this 'controlling' for differences between the groups can only be as good as the data collected. If the two groups differ in terms of unobserved factors, such as previous work history and qualifications, or more hard to measure factors such as motivation, then unless data is available on these factors they cannot be controlled for. However, it is usually a reasonably safe assumption that over fairly short intervals of time eligible populations will differ

only slightly on these factors. If this assumption is correct not being able to control for these factors should not seriously bias the estimate of programme additionality. A general rule might be that if the eligible populations are of similar size and profile (on the observables) in the before and after periods then there is not likely to be a problem. If the size and/or profiles are quite different then there may be a serious problem which controlling for the observables will not address.

### *Historical change*

A more serious threat to before-after designs is that, even if the before-after groups are comparable, change over time might be attributable to general historical change rather than to change brought about by the programme. Historical change might occur for several reasons:

- most outcomes of interest in labour market programmes will change as the general economic climate changes. If the economic climate is different in the 'before' period than in the 'after' period, change will occur irrespective of whether the programme is introduced (NB depending on the nature of the economic change this can have the effect of making the impact of the programme look either bigger or smaller);
- even in periods of economic stability there is likely to be some degree of 'random' or 'unexplained' change between any two periods in time;
- if other programmes are introduced, or changed, at about the time that the programme of interest is introduced the change over time will include the impact of these other programmes as well;
- in some instances new programmes are introduced because a problem within a particular group has been identified. So a programme might be introduced at a period when positive outcomes are known to be particularly low. In many instances, even without a programme intervention, a low rate of positive outcomes will improve over time;
- there may be other trends occurring that introduce change over time, such as increasing activity rates for women.

Unfortunately there are no reliable methods for disentangling historical change from programme change in a basic before-after design where there is just one 'before' set of outcome measures and one 'after' set. Most evaluators who use this design will look for evidence that historical change has either not occurred or that the size of the historical change would be small relative to the change observed – in which case it can be argued that *most* of the observed change is due to the programme. This suggests that a basic before-after design is appropriate in instances where:

- there are no expected historical changes that might affect outcomes (which will rarely be the case for labour market programmes);
- the before-after outcomes are measured over such a short period of time that historical change over the period will be very small. (This is unlikely to be practical for labour market programmes because the full impact of programmes can only be expected after several months at least.);

- the impact of the programme is expected to be large enough to ‘swamp’ historical change. (This might be the case for some mandatory labour market programmes, but is less likely to be true for voluntary programmes, especially if the participation rate is low.)

The designs discussed in Sections 8.2 and 8.3 are attempts to enhance the basic before-after design to avoid some of these difficulties.

### **8.1.3 Before-after designs with voluntary programmes**

As hinted at in the previous section, before-after designs can be particularly inappropriate for voluntary programmes, especially those with low take-up rates. The reason for this is that the impact of these programmes on the whole of the eligible population (as opposed to the sub-set of participants) will usually be very small. In these instances disentangling change due to the programme from historical change is probably impossible.

The problem might be largely solved if the ‘before’ sample (i.e. the comparison sample) could be restricted to would-be participants rather than the whole of the eligible population. In this case change over time would be measured for the group where the programme impact is concentrated – and in this instance the change over time might be large enough to ‘swamp’ any expected historical change. However in most instances it is simply not possible to identify ‘would-be participants’ from the before-period.

The only exception might be in instances where the eligible population is static over time and a cohort approach is appropriate. In this instance it might be feasible to take a sample of participants from the new programme and collect outcome data on them retrospectively from the before period (either using retrospective questions in a survey or using historical administrative data). However it is hard to think of examples that might fit these criteria. Certainly programmes to help people into work are not appropriate because the eligible population for these programmes is not static over time. One possible (fictional) example might be a voluntary programme to help those in stable part-time jobs to increase their hours (perhaps help with childcare for lone parents might be an example). If those in stable part-time jobs are a reasonably static population (a rather doubtful assumption), a sample of participants might be asked retrospectively about the hours worked in the ‘before’ period. Any change in these hours, on average, might then be attributed to the programme. (In practice it would be desirable to compare participants with other similar groups since change in hours as people, and their children, age is a natural ‘historical’ or ‘maturation’ change. This would then become a difference-in-differences design – see Section 8.3.)

## **8.2 Interrupted time-series designs**

The problem of disentangling programme change from historical change in before-after studies can sometimes be overcome by extending the number of before

periods and, if possible, the number of after periods, to give a time-series. If a break in this time series occurs at, or shortly after, the time when the new programme is introduced, this is interpreted as the impact of the programme. Designs based on this principle are known as 'interrupted time-series designs'.

Time series help in ruling out some possible alternative explanations for change. In particular, as long as there are no related programmes introduced at about the same time, then a sudden change in the time series would be fairly conclusive. This is especially true if the change observed with the programme is larger than the change observed between any two of the 'before-periods'.

If it can also be demonstrated that the change coinciding with the introduction of the programme is sustained over time then the evidence is even stronger. This is the argument for using several 'after' periods.

Because of the need for reasonably long series of data in time series analysis, the method is best suited to administrative data analysis, although in some instances it might be possible to make use of large-scale continuous, or near-continuous, surveys such as the Labour Force Survey. The issues around whether to measure time series on cohorts or on fresh cross-sectional samples are the same as for the simple before-after study – if the eligible population stays constant over time the cohort approach is preferable (although not essential). Otherwise fresh cross-sectional samples of the eligible population are appropriate.

### **8.2.1 *The analysis of time series data***

There is a huge statistical and econometric literature on the analysis of time series data. The more sophisticated models allow for seasonality, and for correlations in regression errors. These models could also include terms to adjust for differences in the profiles of the eligible population over time. The main aim of the analysis is to test whether the change observed at the time the programme is introduced is larger than could be expected by chance (given the evidence about 'natural change' from other periods). How sophisticated the modelling needs to be depends upon the complexity of the time series data. In many instances, a time series analysis will incorporate models of the impact of changes in the economy on the outcomes of interest so that these can be 'removed' from the estimates of change.

### **8.2.2 *Problems***

Time series do not completely solve the evaluation problem. If it so happens that the introduction of the programme of interest coincides with other events, or the introduction of other programmes, that have an impact on outcomes then it will still be impossible to prove that the programme of interest is causing the change in outcomes observed.

For programmes that have a delayed or gradual impact the interruption in the time series will occur some time after the programme is introduced. Unless this is anticipated the impact of the programme may simply be missed.

Over and above this, the strong reliance of the final estimates of additionality on the modelling may leave the estimates open to question – small changes in the model might lead to large changes in the estimates. It would be appropriate to thoroughly test the sensitivity of the models to changes in the underlying assumptions.

As with the basic before-after design, time series approaches are most convincing for programmes that have a reasonably large impact, since it will then be fairly easy to detect this impact over and above the ‘background noise’. This means that the approach will often be *unsuitable* for voluntary programmes, especially those with low participation rates.

### **8.3 Difference-in-differences enhancements**

Both the basic before-after design and the interrupted time series design estimate additionality due to a programme by calculating the (adjusted) difference in outcomes between the before and after periods. However, both of the designs can be unconvincing if there is a suspicion that the difference might have arisen because of other events happening at or about the same time the programme was introduced. Two possible ways of enhancing these designs are described in this section.

#### *Non-equivalent control groups*

One way that evaluators have dealt with competing historical change interpretations is to argue that, for programmes targeted at specific groups, ‘natural’ or ‘other cause’ historical change would affect not only the target group but other similar groups as well. If this is the case *extra* change in the target group between the before and after periods, relative to the change in the ‘parallel group’, might be attributed to the programme. The ‘parallel group’ is referred to in the literature as a ‘non-equivalent control group’. The mathematics are essentially trivial: a difference before-after is calculated for the non-equivalent control group (this is the first difference). A second before-after difference is calculated for the target group (this is the second difference); and programme additionality is calculated as the difference between these two differences (i.e. the difference-in-differences).

A prominent example of the approach was the recent evaluation of NDYP. NDYP was targeted at unemployed 18-24 year olds. It was demonstrated by the evaluators (NIESR) that historical change for this group was similar to historical change for the unemployed in the 25-29 age group. So 25-29 years olds were used as the non-equivalent control group for the 18-24 year olds. Since greater change over time in moves off unemployment benefit were observed for the 18-24 year

olds than was observed for the 25-29 year olds, the extra change was attributed to NDYP.

This method is most convincing if it can be demonstrated that outcomes for the target group and the non-equivalent control group have historically moved in parallel. The best source of data for this will usually be administrative records, assuming that outcome data, or a reasonable proxy, are available. An alternative source might be large repeated surveys.

For most labour market programmes the approach has limited applicability because of the difficulty of finding an appropriate non-equivalent control group. The exception is an adaptation of the method used in matched area comparison evaluations. This is covered in Section 8.7.

It is worth noting that there extensions to the approach have been proposed. For instance, some researchers suggest the use of ‘difference-in-difference-in-differences’ approaches which are an attempt to control for different reactions to economic change in the two groups being considered.

#### *Non-equivalent outcome variables*

Another approach that has proved convincing in other contexts but has, to our knowledge, not been used in the evaluation of labour market programmes, is to compare change over time in the outcomes of interest with change in other related outcomes. These *non-equivalent outcome variables* have to satisfy the properties that historically they have changed in parallel with the outcomes of interest, *but* they are not affected by the programme.

A fairly famous example of the method was the evaluation of the compulsory breathalyser tests in Britain in 1967. Two outcome variables were compared: accidents on weekend nights, the rate of which was expected to be changed by the introduction of the tests; and accidents in commuting hours, the rate of which was not expected to be changed. What was observed was a sharp drop in the first rate and little or no change in the latter. This provided strong evidence that the tests had had an impact.

It is not easy to see how the method could be applied to the evaluation of labour market policies. It is unlikely that a very convincing non-equivalent outcome variable could ever be identified, so it is not to be expected that any evaluation will ever rely solely on this method. It might however be used in some contexts as a double-check on the estimates of additionality made using other methods.

#### **8.4 Other time-series approaches**

Three other time-series approaches to estimating additionality are worth mentioning.

### *Removed treatment design*

In exceptional circumstances programmes are introduced and then removed. In these instances time series would be expected to reveal two 'interruptions': one at the time the programme is introduced and one at the time it is removed. One would expect the change in the time series with the introduction of the time series to be largely reversed after the programme is removed.

This approach could, in principle, prove very powerful. It is, however, rarely, if ever, used in government programme evaluations for the simple reason that if a policy is abandoned there is usually very little interest in doing a post-mortem.

### *Delayed treatment design*

In some instances a programme is piloted within a few areas before being introduced nationally. If time series data show a before-after change in outcomes in the pilot areas that is then replicated in other areas when the programme is introduced nationally, this would provide very strong evidence of the impact of the programme.

### *Regression discontinuity design*

If eligibility for a programme is based on a scoring system, with only those with a score above (or below) some threshold being eligible for the programme, a difference-in-differences approach can sometimes be adopted, the control group being those with an ineligible score.

## **8.5 One-to-one matched comparison group design**

As was noted earlier, time series approaches to evaluation tend to be unconvincing for voluntary programmes with low participation rates simply because, measured across the whole of the eligible population, the impact of the programmes tends to be very small. For example, if one in ten participants finds work specifically because of a programme (so that additionality is 10% for this group), if just 5% of the eligible population participate, this translates to additionality of just 0.5% for the whole of the eligible population. Time series analyses will rarely be able to detect impacts of this size.

Because of these difficulties different quasi-experimental designs and analyses have been developed that are useful in exactly these circumstances. There are broadly two approaches: the one-to-one matched comparison design that is described in this section and the more general approach which we've labelled 'statistical modelling of existing data' since it covers a range of analysis methods rather than an alternative quasi-experimental design as such. The former is typically used when data on outcomes is to be collected by survey, in which case the size of the programme and comparison groups have to be kept reasonably small. The 'statistical modelling' methods are used when data on outcomes is readily available, usually from administrative records or from surveys carried out for other purposes. The one-to-one matched comparison group design is described in this section.

The one-to-one matched comparison group design makes use of the fact that additionality in voluntary programmes is concentrated within the programme participants, and not spread across the whole of the eligible population. In fact it is specifically assumed that the programme has no impact on non-participants, so that a reasonable estimate of the counterfactual can be estimated from this group.

As with all the quasi-experimental methods, two groups are compared: a programme group and a comparison group. In the matched comparison design the programme group is selected from the pool of participants. The comparison group is selected from the pool of non-participants. If it is assumed that outcome data is expensive to collect, which would be the case if it was to be collected by survey, the comparison group will typically be chosen to be the same size as the programme group (hence a *one-to-one* matched comparison), although one-to-two or one-to-n matches are also possible.

Matched comparison group designs are controversial because of the way the comparison group is selected. For a comparison group to give an accurate estimate of the counterfactual, the group must have the same profile of characteristics as the programme group, insofar as these characteristics are related to outcomes, with the single exception that they are not exposed (or in this case *choose* not to be exposed) to the programme. But, unlike mandatory programmes, in this case the programme group are self-selecting and in large part the reasons for participation will not be fully known or understood. This means that constructing a suitable comparison group is very difficult.

For example, those volunteering for a New Deal labour market programme might tend to be people who are highly motivated to find work, have struggled to find suitable work in the past, and who have lower qualifications than non-volunteers. If this is the case the comparison group should be made up of a similar profile of people, that is those who are, highly motivated, have struggled to find work in the past, and with lower qualifications. And if instead, the programme group over-represents motivated persons whose previous experience makes finding work easier than average, the comparison group should be made up of a similar mix of people.

Any mis-match between the programme and comparison groups will tend to bias the estimates of additionality derived, sometimes severely. And the biggest problem with the design is that the bias will tend to go unobserved. For example, if participants *are* more highly motivated to find work than non-participants but no data on motivation is available, it is probable that the programme and comparison groups selected will differ quite considerably on this one factor. And if in addition, highly motivated persons are more likely to experience a positive outcome irrespective of whether they are in the programme or comparison group, those in the programme group will have better outcomes than those in the comparison group even if the programme has no impact whatsoever. So, in this example, the programme would appear to have an impact even if in reality it does not. But,

because no data is available on motivation, the imbalance between the programme and comparison groups will go unnoticed.

To avoid bias in a matched comparison group design the following two criteria need to be met:

- evaluators must have a complete understanding of the factors that influence participation (in particular it is important to know about factors that influence both participation and outcomes)
- the data on these factors must be available for all, or a sub-set, of both participants and non-participants.

Typically the understanding about which factors influence participation will come from a combination of theory and specially commissioned research.

*If* the two criteria given above can be met the matched comparison group design will be a robust evaluation method.

The steps to be followed are:

- (i) collect data, if necessary, on the factors that influence participation;
- (ii) select a programme group from the pool of participants;
- (iii) match each programme group member to a 'similar' non-participant to form the comparison group;
- (iv) collect data on outcomes for both the programme and comparison group members.

Each of these steps is discussed in turn below.

### **8.5.1 Collecting data on the factors that influence participation**

In exceptional circumstances theoretical considerations may suggest that all the factors that influence participation are available on administrative systems. However, in most instances, a special data collection exercise will be needed to collect the data.

The ideal is:

- The data should be collected before participation, because the act of participation may itself change some of the factors. For instance if motivation affects participation, but participation increases motivation, it is the pre-participation motivation that is important, since it is this that contributes to the counterfactual;
- The data should be collected *just* before participation if at all possible, to avoid the chance that intervening events will have an influence on participation.
- The data should be collected on many more non-participants than participants to provide an adequate pool from which the comparison group might be drawn.

The way this was handled on the evaluation of the National New Deal for Lone Parents (NDLP), which uses a one-to-one matched comparison group design, was to send a postal questionnaire to a large sample of eligible lone parents before participation. This questionnaire was designed to collect data on all the *expected* factors influencing participation. The programme group was then defined as those lone parents returning a postal questionnaire *and* participating in the programme within a reasonably short interval of time. The comparison group was drawn from the large pool of lone parents who sent a questionnaire back but who did not participate in the programme before the date when outcome data was collected.

### **8.5.2 Selecting the programme group**

The selection of the programme group for the evaluation will tend to be fairly straightforward. Ideally, as was noted in the previous section, it will be selected from the pool of participants for whom *recent* data on the factors influencing participation is known.

As with all evaluations, there may be arguments for over-sampling or over-representing sub-groups for whom separate estimates of additionality are required.

### **8.5.3 Selecting the comparison group/propensity score matching**

To create the comparison group each member of the programme group is uniquely matched to a non-participant. The aim of the matching is to create a comparison group that is the same, to within an acceptable margin, as the programme group on all the factors that influence participation.

In early versions of the matched comparison group design the number of factors to be used in the matching would typically be quite small (primarily because they were simply the factors available from administrative systems). To do the matching with just a few factors the usual approach would be to divide the programme group and the pool of non-participants into 'cells' based on the factors, so that all the members of the cell had the same characteristics. Each programme member within a cell would then be 'matched' to a non-participant by selecting a non-participant from within the same cell at random.

This method works well if the number of factors to be matched on is small so that the number of cells is also fairly small. But, as the number of factors increases so does the number of cells, and the method quickly becomes very problematic. For instance, it would not be uncommon with only a fairly small number of factors for cells to be created with participant members but no non-participants in the same cell to match to. In these instances rules have to be set up to allow for matches to be made from 'close' cells. Without a good theoretical basis for these 'rules' the matching starts to look arbitrary.

In recent years an alternative method of matching has been developed based on propensity scores. The idea is broadly as follows:

- An estimate of the probability (propensity) of participating is calculated for all participants and non-participants. This will usually be done by fitting a logistic or probit regression model to the data, using the factors thought to influence participation as the predictors.
- Each member of the programme group is then matched to the non-participant with the closest propensity score.

By matching in this way the comparison group is guaranteed to have a very similar profile of propensity scores to the programme group. This means that, if for instance, the programme group disproportionately includes people with a high propensity to participate because they are highly motivated, but need help with finding work, then they will be matched to non-participants with a similar propensity to participate (although in fact they happened to choose not to) who will also be highly motivated but who need help in finding work. The great advantage of the propensity score method is that it reduces the matching task to a univariate matching problem rather a multivariate problem. Propensity scores are to be the subject of a separate paper in the Working Paper series.

There are a number of issues/problems that may have to be addressed before the comparison group is created:

- a decision is needed as to how close the nearest non-participant to a participant has to be before a match is made;
- once this decision is made, there may be some members of the programme group for whom a suitable match simply cannot be found (this is known as the support problem). For programmes with low participation rates, where there is a large pool of non-participants to choose matches from, this is unlikely to be a major problem;
- in some instances one non-participant may be the closest match to two or more members of the programme group. In such cases a decision is needed on whether the non-participant can be used more than once.

These issues are currently being worked through for the evaluation of the national NDLP programme. It is probably too soon to draw general conclusions at this point.

#### **8.5.4 Problems in the interpretation of matched comparison designs**

It should be noted that the method of propensity score matching is only as good as the data collected. If the propensity scores calculated are inaccurate, either because not all of the relevant data on factors affecting participation is available, or because some of the factors are measured inaccurately<sup>3</sup>, the matching will not be perfect.

---

<sup>3</sup> Most surveys can only collect indicative data on latent factors such as motivation. The inability to measure such important factors accurately means that there will be residual differences in, say, the motivation of the programme and comparison groups even after matching.

Typically this will mean that there are residual, uncontrolled, differences between the programme and comparison groups. These differences are quite likely to bias the estimates of additionality, although it will be impossible in most cases to even approximately estimate that bias. One of the main difficulties with the matched comparison approach is that there is no test of whether or not the matching is adequate to control bias.

#### **8.5.5 *Collecting data on the outcomes from the programme and comparison groups***

One-to-one matched comparison group designs are typically used when outcome data is to be collected by survey. Non-response to these surveys can create particular theoretical difficulties in a matched comparison setting because matched pairs will typically not choose to respond, or not-respond, in pairs. So at the analysis stage some of the balance in the propensity scores may be lost. Again, these issues are currently being thought through for the NDLP evaluation and it would be premature to draw conclusions at this point, but one possible way forward would be to 're-match' the final samples. This would mean reconstructing the comparison group for the responders in the programme groups, and discarding any respondents for whom a suitable match cannot be found.

#### **8.5.6 *When matched comparison designs are appropriate***

The matched comparison design approach will give a reasonably robust estimate of additionality if:

- all the factors affecting participation are known
- data on all these factors can be collected.

Arguably, the lower the participation rate the more reason there may be for thinking participation in individual cases is for very personalised reasons. In these instances it may be wrong to assume that data on the reasons for participation can be collected thoroughly and accurately. If this is correct, the method may not be appropriate for programmes with very low participation rates. Unfortunately these are the very circumstances under which the alternative quasi-experimental methods also fail. In these instances a randomised trial of participants may be the only option.

#### **8.5.7 *Matched comparison group designs used in combination with other quasi-experimental methods***

A variation on the matched comparison group design can be used in before-after studies or in matched area comparison designs for either mandatory or voluntary programmes (see Section 8.7). In these instances the programme group is usually selected from the whole of the eligible population at a point in time when, or in areas where, the programme has been introduced. This programme group is then matched to members of the eligible population from the 'before' period or from areas where the programme is not being piloted. Since the eligible population is not self-selecting the matching procedure is less controversial (i.e. there is no requirement to identify and collect data on the factors influencing participation).

Another variation, *specifically for voluntary programmes*, would be to select the programme group from the population of voluntary participants but then to match these to members of the eligible population either from a period before the programme was implemented (i.e. a before-after approach) or from areas where the programme has not yet been introduced (a matched area approach). In both instances data on the factors thought to affect participation would have to be collected (so the timing of the data collection for a 'before-after' approach may rule this option out) but since the pool of non-participants from which the comparison group will be selected will include people who would participate if given the opportunity, the match between the programme and comparison groups will arguably be better than in the standard matched comparison group design. Nevertheless, the benefits of doing this have to be weighed against the disadvantage of having a comparison group selected from a different period or from different areas to the programme group.

## **8.6 Statistical modelling of existing data to evaluate voluntary programmes**

The one-to-one matched comparison group design creates a comparison group that is about the same size as the programme group, but which is usually selected from a much larger pool of non-participants. This is appropriate if collecting data on outcomes is expensive, since it is then wasteful to select a larger comparison group than is strictly necessary. If, however, outcome data can be collected very cheaply or at no expense, greater precision can be gained by making use of more of the non-participant group.

The econometrics literature suggests several methods for making estimates of the counterfactual under these circumstances, all of which are rather technical. A full description is not given here, but a very brief description of the three main methods is given: propensity score matching with kernel weighting, the instrumental variables estimator, and the Heckman selection estimator.

### **8.6.1 Propensity score matching with kernel weighting**

As with the matched comparison group method, a propensity score is estimated for all participants and non-participants in a programme. But instead of using this score to find a *single* match for each participant, the counterfactual is estimated *per participant* as the weighted sum of outcomes across all non-participants. The size of the weights (which in total sum to 1) per non-participant depends upon the distance between the propensity score for the participant and the non-participant, small distances getting the greatest weight. The weights are known as kernel weights. The actual distribution of the weights is most commonly chosen to follow a normal distribution – the weight each non-participant gets then depends on the variance set for this normal distribution. As in the one-to-one matched comparison, for this approach to give an unbiased estimate of the programme

impact the propensity score has to be free of any errors that are related to outcomes.

### **8.6.2 *The instrumental variables estimator***

*If a variable can be identified that is related to, or correlated with, participation but is not correlated with any other variables that are related to outcomes (including those not observed), then comparing outcomes across different values of this instrumental variable allows for the participation rate to be varied whilst holding all other factors constant. This gives information about how outcomes relate to participation, which in turn allows for an estimate of additionality to be made. In practice it is very difficult to find an appropriate instrumental variable.*

### **8.6.3 *The Heckman selection estimator***

The key difficulty with the instrumental variable approach is that it requires an instrument that is uncorrelated with the errors in the outcome model (which means it must be uncorrelated with unobserved or unmeasured variables that nevertheless impact on outcomes). This is generally unachievable in non-experimental settings. Heckman proposed an alternative approach where, instead of assuming no relationship between the errors in the outcome equation and the instrument to be used (the instrument is typically a propensity score), he suggests how the relationship between the two might be modelled (and hence controlled for). The main difficulty is that very strong (and usually untestable) assumptions about the form of the relationship have to be met for the method to give unbiased estimates of programme impact.

## **8.7 Matched area comparison design**

All of the quasi-experimental methods discussed so far have been appropriate for the evaluation of programmes that are introduced nationally at one point in time, although all of the methods can also be used to evaluate pilot programmes. However, the design discussed in this section, the matched area comparison, can only be used to measure additionality when a programme is piloted within some parts of the country.

The basic design for a matched area comparison is as follows:

- (i) A small number of areas (typically this has been less than 10) are selected within which the new programme will be piloted. The programme group is selected from these areas.
- (ii) These areas are matched to a number of areas with similar characteristics where the programme is not being piloted. The comparison group is selected from these areas. In a variant of the basic matched area design, a comparison group is selected from the rest of the country rather than from matched areas.
- (iii) Outcomes are then collected on both the programme and control groups.

The first two of these steps are discussed in turn below.

### **8.7.1 *Selecting the pilot areas***

Pilot areas are generally selected purposively rather than randomly. The criteria for selection will vary from programme to programme, but the areas will typically be selected to cover a range of labour market conditions. The area unit chosen will depend upon the programme – in most recent labour market programme evaluations it has been an Employment Service area.

The number of areas selected tends to be kept small for practical reasons. The small number does, however, have some implications for statistical inference: variation in additionality between areas cannot realistically be included into the standard error estimates (so standard errors tend to be under-estimated), and, if the estimates of additionality do vary from area to area, it is difficult to state what would be the average estimate of additionality if the programme was to be introduced nationally. Defenders of the approach would argue that the pilot is designed to test whether or not the programme ‘works’ under the conditions imposed – inference to the national picture is a secondary issue.

It is worth noting at this point that if inference to the national picture was a primary aim, randomisation of areas, as discussed in Section 7.5, would be a better approach. This would however be a much more expensive option as it would inevitably mean that the programme had to be piloted in more areas.

### **8.7.2 *Selecting the comparison areas***

Under the standard matched area comparison design, the pilot areas are matched, usually one-to-one but not always, with areas that have similar characteristics. No consensus seems to have developed about how this matching should be done, but most commonly it will be done on a small set of economic indicators such as the unemployment rate. In practice, once sample sizes, competing initiatives, and willingness of areas to take part have been taken into account, the choice of pilot and matched areas may be rather constrained.

One way to approach the matching, which borrows heavily on the difference-in-differences theory (see section 8.3), is that pilot areas should be matched to non-programme areas where the eligible population have, over a series of years, experienced the same, or, parallel changes in the outcome measures of interest. A divergence in these outcomes once the programme is introduced in the pilot areas would then provide strong evidence for a programme effect. This method, of course, depends upon there being historical data available on outcomes for the eligible population. In most instances this data will not be available, but closely correlated data may be. For instance, for a programme aimed at moving the unemployed into work, historical data on moves off benefit would probably be a good proxy.

One problem that was encountered on the evaluation of the NDLP prototype was that the pilot areas and the comparison areas, which matched well at the time they were selected, had drifted apart by the time outcome data was to be collected.

Since it is very difficult, if not impossible, to control for this in the analysis, an alternative is to draw a comparison sample 'nationally' (i.e. across all non-pilot areas) rather than from matched areas. This means that the comparison sample will cover a wide range of local economic conditions, and there should always be at least a sub-sample of the comparison sample that is from areas similar, at the time of analysis, to the programme sample. Furthermore, by taking a national comparison sample rather than an area comparison sample, the national sample can be used to describe the eligible population (an analysis that is not strictly necessary for the estimate of additionality, but which can be very useful for planning for a national extension of the programme). Nevertheless, taking a national comparison sample can lead to the situation where much of the data collected on the 'comparison sample' is simply thrown away because it cannot be made comparable to the population in the pilot areas. *If* the matched areas are selected well and there is no divergence over time, the matched area approach makes best use of the comparison sample.

A further point to note. By hand-picking the comparison areas evaluators might be accused of trying to ensure a good measure of additionality is achieved. (This might be done by always picking matched areas with *slightly* more depressed labour market conditions than the pilot areas.) Any such accusations could be partially avoided (although there will always be those who doubt the procedures) if areas were selected in pairs and one of the pairs was randomly assigned to the programme.

### **8.7.3 Problems in the interpretation of matched area comparison designs**

The interpretation of matched area comparison designs can prove problematic because observed differences between the programme and comparison groups could be attributable to three things:

1. Difference attributable to the programme;
2. Differences in the profiles of the programme and comparison groups;
3. Differences in the labour market conditions between areas.

The first of these differences is the difference we would like to measure.

The second difference is potentially a problem, but to some extent at least, can be dealt with by controlling for differences between the groups at the analysis stage, either by regression analyses, weighting the samples to a common profile, or by matching individuals within the programme group to a similar person in the comparison group (i.e. created a matched comparison group). As was noted for before-after comparisons, this 'controlling' for differences can only be as good as the data collected. If the groups differ on important factors that no data is available on, then the difference between the groups in terms of this factor is likely to remain even after controlling for 'observed' differences.

The greatest threat to the additionality estimates is that there may simply be systematic differences between the pilot and comparison areas that account for

some, if not all, of the differences in outcomes. This should not be a major problem if the programme impact is expected to be large – with careful matching of areas the difference in outcomes due to ‘area effects’ is likely to be reasonably small, so with a large difference between the groups it will be possible to state that most of the observed difference is due to the programme. However, between-area differences can be a major problem if the impact is expected to be small, as is the case with many voluntary programmes.

#### **8.7.4 A difference-in-differences enhancement**

One way to strengthen the analysis is to use a difference-in-differences approach, as introduced in Section 8.3. We have already described above how historical data on outcomes can be used in the matching. But, in addition, additionality estimates can be made more precise if, instead of taking a straightforward difference between the programme and comparison groups, additionality is measured as the before-after difference in the pilot areas minus the before-after difference in the comparison areas. This is equivalent to controlling for differences at baseline (because the same estimator can be written as the difference between areas after the pilot has been introduced minus the difference between areas ‘before’ or at baseline). Nevertheless, even with this enhancement, it is still likely that a matched area comparison will be inconclusive for programmes with small expected impacts.

#### **8.7.5 A possible improvement for voluntary programmes**

The reason matched area comparisons are likely to fail for voluntary programmes is because there is no means of identifying ‘would-be’ participants in the comparison areas. This means that, under the standard design, the whole of the eligible population, rather than simply the participants, have to be compared.

We described in Section 8.5.6 how a matched comparison design might be combined with a matched area comparison. This method however relies on the ability to identify a comparison sample who are like the participant sample in the pilot areas in all relevant ways, and as should be clear from Section 8.5, this is not an easy task.

An alternative method, a version of which was used on the evaluation of the pilot New Deal for Disabled People (NDDP) programme, is to narrow down the eligible population within the pilot and comparison areas to a group from which most participants will be drawn. In NDDP this was defined as those disabled people identified in a survey as being ‘close to the labour market’. If participation rates within this sub-group are much higher than the overall participation rates then additionality for this group should be much higher than average. So by comparing a programme group comprising those ‘close to the labour market’ with a comparison group also ‘close to the labour market’, there is a far better chance of being able to detect additionality (at least for this important sub-group). Furthermore, because the additionality estimate should be reasonably large,

assuming the programme does 'work', the estimate of additionality should be large enough to be able to claim that most of it is a programme effect rather than a between-area effect. Nevertheless, the sample sizes in the NDDP pilot were too small to detect significant additionality and no thorough assessment of the method was possible. For the method to be useful it would have to be demonstrated that:

- most participants were indeed from the population of those 'close to the labour market' (if this criterion is not met the design will fail because no estimates of additionality will be available for those not close to the labour market)
- the participation rate amongst those 'close to the labour market' is high enough for reasonably large additionality figures to be expected.



## 9 SAMPLE SIZE CALCULATIONS

This section gives formulae and examples of how to calculate sample sizes for the designs described in this report.

### 9.1 Sample size calculations for randomised trials

To estimate the sample size needed for a randomised trial several factors must either be estimated or decided upon.

To be decided upon:

1. The smallest difference (i.e. impact) that it is important to be able to detect (i.e. the smallest difference for which a statistically significant result is wanted)	= d
2. The proportion of the total sample that will be allocated to the programme group. <i>Usually a=0.5.</i>	= a
3. The significance level to be used in the statistical tests. <i>Usually a is set at 5%. For the level of a chosen, tables of the Normal distribution are needed to estimate <math>z_{a/2}</math> (assuming a two-sided test is to be used). If a =5%, <math>z_{a/2}=1.96</math>.</i>	= a
4. The power of the study (i.e. the probability of finding a significant difference between the programme and control groups when a real difference of d exists). <i>Usually the power is set at either 80% or 90%. For the power chosen, tables of the normal distribution are needed to calculate <math>z_b</math>. If the power=80%, <math>z_b=0.824</math>; if the power is 90%, <math>z_b=1.282</math>.</i>	= 1 - b

To be estimated:

5. For a binary outcome, the approximate proportion of positive outcomes in the control group; For a continuous outcome, the approximate variance of the outcomes in the control group	= p = s <sup>2</sup>
6. The proportion of each group that it will be possible to collect outcome data on (i.e. the response rate).	= r

Then, for a binary outcome, the total number (n) to be entered into the trial is:

$$n = \frac{1}{r} \left\{ \frac{z_{a/2} \sqrt{\frac{p(1-p)}{a(1-a)}} + z_b \sqrt{\frac{p_1(1-p_1)}{a} + \frac{p_2(1-p_2)}{(1-a)}}}{p_1 - p_2} \right\}^2$$

where  $p_1 = p + d; p_2 = p; p = ap_1 + (1 - a)p_2$ .

And for a continuous outcome, the total number (n) to be entered into the trial is:

$$n = \frac{1}{r} \left\{ \frac{z_{a/2} + z_b}{d} \right\}^2 \frac{s^2}{a(1-a)}.$$

If separate estimates for sub-groups are needed then the calculations need to be repeated for each sub-group, and the total trial size set accordingly.

**Example: Sample size for a randomised trial with (a) randomisation of the eligible population, and (b) randomisation of those participating.**

Suppose in a trial of a voluntary programme we wish to be able to detect additionality of 5% amongst volunteers. But only 10% of the eligible population volunteer, so that additionality measured across the whole of the eligible population is just 0.5%.

Assume that allocation will be in the ratio 50:50. The significance level to be used is 5%, and the required power is 90%.

The outcome variable is binary. In a trial of the eligible population approximately 10% of the control group will experience positive outcomes; in a trial of participants approximately 30% of the control group will experience positive outcomes.

Outcomes are to be collected by survey. The expected response rate is 75%.

So, for the randomised trial of the eligible population,  $d=0.005; a=0.5, p=0.1,$  and  $r=0.75$ . The sample size needed in this instance is an enormous 207,000.

And for the randomised trial of participants,  $d=0.05; a=0.5, p=0.3,$  and  $r=0.75$ . The sample size needed in this instance is just 4,900.

Note that although the participation rate is 10%, the sample size for the trial of participants is less than 10% of the sample size needed for a trial of the eligible population. This occurs because in the trial of participants random differences between the programme and control groups for non-participants do not have to be allowed for.

Deciding on the sample size parameters for a randomised trial is no easy matter and needs considerable thought. A few of the issues are outlined below:

*1. The significance level*

The default for statistical tests is to use a 5% significance level. This means there is a 5% probability of concluding that a new programme is significantly different to the old when in reality there is no real difference. For many types of trial, setting this probability as low as 5% is appropriate because it would be very undesirable to conclude a new programme or treatment was better than the old if in fact it isn't. Arguably for labour market programmes however, the new programme should be given the benefit of the doubt, in which case a significance level of, say, 10% might be more appropriate.

*2. 1 or 2 sided tests*

If it is absolutely clear that a new programme *cannot* be worse than the existing programme then a one-sided test can be applied rather than a two sided test. This will reduce the sample numbers required. In practice it will be impossible for the positive or neutral effect of the new programme to be known with such certainty so two-tailed tests remain the norm.

*3. Statistical power*

It is usual to set the power of a trial to either 80% or 90%. With a power of 80% this means there is a 20% chance that no significant programme impact will be detected even if the true difference is  $d$ . Since we have suggested above that the significance level for labour market evaluations could be set at greater than 10% if new programmes were to be given the benefit of the doubt, for the same reasons we would suggest that the power of the trial should be set at 90% or higher.

*4. Statistical tests v. fixed confidence intervals*

The sample size estimators set out above assume that the main objective is to detect programme impacts that are 'significantly' different to zero. However some thought needs to be given to whether or not this is what is of interest. An alternative approach would be to set the sample size so that the confidence interval around any estimate of additionality will be no greater than some fixed amount. This might, for instance, be the best approach from a cost-benefit perspective.

*5. Allowing for multiple testing*

It is considered good practice to make estimates of additionality for a programme using as many data sets, and as many experimental and quasi-experimental estimators, as possible. Then if the various methods all give similar results the findings can be considered as 'robust'. The danger with this approach is that with multiple tests and approaches the probability of finding at least one 'significant difference' increases quite alarmingly. There are formal methods for adjusting significance levels to allow for multiple testing, but irrespective of whether these formal methods are used, we would suggest that at a minimum any single estimate that differs greatly from others be treated with considerable caution.

## 9.2 Sample size calculations for randomised area trials

For a randomised trial of areas the sample size calculations need to take into account the fact that the standard errors of the outcomes for the programme and control areas will include an additional ‘between-area’ component of variance. A crude inflation factor can be applied to the sample size calculations of Section 9.1 to account for this, namely:

$$n_a = n(1 + (m - 1)r)$$

where  $n_a$  = sample size of individuals for a randomised trial of areas;

$n$  = sample size of individuals for a randomised trial of individuals (calculated using the formulae of section 9.1);

$m$  = average sample size per area; and

$r$  = the intra-cluster correlation coefficient.

$r$  is a measure of the extent to which areas differ from one another in outcomes, and is calculated as:

$$r = \frac{s_b^2}{s_b^2 + s_w^2}$$

where  $s_b^2$  is the variation between areas and  $s_w^2$  is the variation within areas.

To calculate the sample size for a randomised trial of areas an estimate of  $r$  will be needed. Often it will be possible to make this estimate using historical administrative data and, if necessary, using a proxy outcome measure (such as the percentage of the eligible population leaving benefits as a proxy for the percentage entering work). To make an estimate of  $r$  in the instance where outcomes are expressed as a percentage the following steps will suffice (the percentage leaving benefits being used as an illustration):

- (i) calculate the percentage leaving benefits in each area;
- (ii) calculate the variance in this percentage across areas (this equals  $s_b^2 + \frac{s_w^2}{N}$  where  $N$  is the average number per area on which the percentages are based – i.e. the average of the denominators);
- (iii) calculate the overall percentage,  $p$ , and estimate the total variance in outcomes (this equals  $p(100 - p)$  which can then be set equal to  $s_b^2 + s_w^2$ );
- (iv) solve for  $s_b^2$  and  $s_w^2$  and estimate  $r$ .

In practice, if randomisation is done within paired areas the sample size can be reduced quite considerably (assuming the pairing is done well). This is because the pairing controls for a proportion of the between-area variance. In this instance the sample size is estimated as:

$$n_a = n(1 + (m - 1)r)(1 - r_p)$$

where  $r_p$  is the estimated correlation in outcomes for matched pairs of areas.

### **9.3 Sample size calculations for the main quasi-experimental designs**

Sample size calculations for the quasi-experimental designs are usually made using the formulae for randomised trials of individuals of Section 9.1. It is sensible however to select somewhat more than the trial calculations suggest in the comparison group to allow for the fact that the data from the comparison group may need to be 'adjusted' (either by weighting, matching, or by statistical modelling) to make it comparable with the programme group.



## 10 EXAMPLES OF THE MAIN EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS

<b>Randomised trial</b>	
Programme	<b>The Restart programme</b>
Programme description	The Restart programme was introduced in 1986 and became a national programme in 1987. The programme involved a six-monthly interview with all unemployed benefit claimants.
Evaluation details	<p>A national sample of 8,189 people approaching six months unemployment was selected. Of these 528 were randomly assigned to a control group all of whom were excluded from the first six-monthly Restart interview but who attended subsequent interviews as appropriate.</p> <p>The randomisation was not strictly adhered to as controls wishing a first Restart interview were allowed one. And in practice one-quarter of the control sample had a Restart interview, although the reasons for this remain unclear. No bias was detected however.</p> <p>Data on outcomes (primarily, changes in benefit status) was collected by survey prior to the second Restart interview. The analysis of the trial suggested that Restart achieved a five per cent reduction in the time spent on benefit.</p>

<b>Before - After Design</b>	
Programme	<b>The introduction of Jobseeker's Allowance (JSA)</b>
Programme description	JSA, introduced in 1996, was a new benefit for unemployed people seeking work that sought to better ensure that jobseekers were meeting their obligations to actively seek and be available for work
Evaluation details	To evaluate the introduction of JSA two samples of benefit claimants were selected, one before the introduction of JSA and one after. Each of the two samples was selected so as to be representative of the unemployed claimants in Britain at the time of selection. The sample members were all interviewed twice: at the time they were selected and after six months when data on outcomes was collected, the primary outcome being entry into paid work. The difference in outcomes between the two samples was interpreted as the impact of the programme. In practice interpreting the change in this way was problematic because of macro economic change that occurred over the interval.

<b>Time series with difference-in-difference</b>	
Programme	<b>New Deal for Young People (NDYP)</b>
Programme description	NDYP is a compulsory programme for those aged 18-24 who have been claiming JSA for 6 months (some other groups can enter early). It aims to improve the employability of participants, and help them to move into employment, through jobsearch assistance (Gateway) followed by participation in an Option for those who have not found work after 4 months. There is further follow-through activity for those who are still unemployed after leaving their Option.
Evaluation details	<p>A key strand of the evaluation of NDYP was an estimate of the impact of the programme using a time series with difference-in-differences design. Changes in the flows on and off unemployment for the client group (18-24s, 6 months plus unemployed) over the period when the programme was introduced were compared with the change in flows for older 6 months unemployed claimants over the same period. Both 30-39 year olds and 25-29 year olds were used as comparator groups in the analysis, which partly reflects the uncertainties involved in deciding on the most appropriate comparison group.</p> <p>Because NDYP is a compulsory programme the impact of the programme was expected to be large, and the difference-in-differences approach did detect a significant NDYP effect. However the estimates of the impact were found to be very sensitive to small changes in macro-economic covariates in the time series models.</p>

<b>One-to-one matched comparison group</b>	
Programme	<b>National New Deal for Lone Parents (NDLP)</b>
Programme description	New Deal for Lone Parents is a voluntary programme which aims to encourage lone parents on Income Support to improve their prospects and living standards by improving their job readiness and by taking up and increasing paid work.
Evaluation details	The impact of NDLP is to be measured using a one-to-one matched comparison design. The data used to match participants to non-participants was collected in a postal survey of a random sample of 70,000 lone parents on Income Support and eligible to join NDLP. The survey was carried out before the 70,000 had had contact with NDLP. Approximately 40,000 lone parents completed and returned a questionnaire.

	<p>These 40,000 respondents were then tracked to see who participated in NDLP and who didn't. Those who participated formed the 'participant group'. These participants were then matched, one-to-one, to non-participants using propensity score matching, the propensity scores being modelled using data from both administrative records and the postal survey responses. The participant and matched non-participants will be interviewed face-to-face in late 2001 to collect outcome data.</p> <p>The success of the approach is largely dependent upon the propensity scores being accurate. Although the postal survey data provides far richer data on the predictors of participation than administrative data alone, it is impossible to be certain that all the factors of importance are captured (and hence controlled for in the matching).</p>
--	---

<b>Matched area comparison (1)</b>	
<b>Programme</b>	<b>Earnings Top Up (ETU)</b>
Programme description	ETU was introduced in October 1996 as a three year pilot scheme. It provided a wage subsidy to people in low-paid jobs working 16 hours or more per week with no dependent children.
Evaluation details	<p>The evaluation used a matched area design. Nine areas were selected. Two variations of the scheme were piloted - Scheme A providing a smaller wage top-up and Scheme B providing a larger wage Top-up. Scheme A, B and a control area were each run in a separate large urban area, a seaside town and a rural area.</p> <p>The evaluation was designed to measure impact on a large number of outcome variables; speed of movement into work; numbers of claimant unemployed; employment duration; employment volumes; wages earned; access and take-up rate among eligible population; effects on eligible populations income; and effects on claiming behaviour. Outcomes were compared during the pilot period between Scheme A, B and control areas and during the pilot period compared with the period pre-ETU (i.e. a matched area comparison with a difference-in-differences enhancement).</p> <p>As with many evaluations with fairly long timescales the evaluation of ETU was effected by changes outside of the control of the evaluators and which made the interpretation</p>

	<p>of findings difficult. For instance the change in administration in May 1997 effected the support for the scheme. ETU had no clear place in the raft of new measures introduced by the new Labour Government. In addition, Jobseeker's Allowance was introduced on the same day as ETU which affected unemployment rates. And finally, the national minimum wage was introduced mid-evaluation taking a large proportion of the ETU eligible population out of entitlement.</p> <p>Over and above these problems the area comparison itself was problematic, with the scheme take-up rate differing greatly across areas. With a small number of areas this makes inference about the estimated impact very difficult. In addition one of the seaside towns based close to London had higher than average incomes so had a smaller pool of entitled population than the other areas.</p>
--	---

<b>Matched area comparison (2)</b>	
<b>Programme</b>	<b>The ONE service pilots</b>
Programme description	ONE is a joint Benefits Agency/Employment Service/Local Authority service providing work focused help and support to clients of working age who are beginning a new claim for benefit.
Evaluation details	<p>The ONE evaluation involves comparing the effectiveness of three methods of delivery. These are the Basic model where delivery occurs through benefit offices, the Call Centre model where clients initiate contact by giving their details to a Call Centre operative, and the Private and Voluntary model where the private or voluntary sector is responsible for delivering ONE. Within each model the impact on the three main client groups, jobseekers, lone parents, and the sick and disabled, is evaluated.</p> <p>The impact of ONE will be assessed by measurement of movement into work. The client survey in pilot and control areas measures this at two stages following participation in the service (4-5 months and 9 months). The cost-benefit analysis uses survey and administrative data on the rate at which people leave benefit to measure the additional employment of ONE. Additionality is measured using a difference-in-differences approach to identify any additional labour market improvement in pilot areas relative to the control areas since April 2000.</p>

## 10 BIBLIOGRAPHY

The following is a short list of papers giving overviews of the main evaluation methods discussed in this paper.

Cook, T. and Campbell, D (1979) *Quasi-Experimentation*. Houghton Mifflin.

Heckman, J and Smith, J.A. (1996) '*Experimental and Nonexperimental Evaluation*' in *International Handbook of Labour Market Policy and Evaluation*, Edward Elgar Publishing.

Purdon, S., Lessof, C., Woodfield, K and Bryson, C. (2001) *Research Methods for Policy Evaluation*. Department for Work and Pensions Working Paper No. 2

Rossi, P., Freeman, H., and Lipsey, M. (1999) *Evaluation. A Systematic Approach. Sixth edition*. Sage Publications.

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