RESEARCH METHODS FOR POLICY EVALUATION

Department for Work and Pensions
Research Working Paper No 2

By

Susan Purdon, Carli Lessof, Kandy Woodfield and Caroline Bryson
National Centre for Social Research

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

First Published 2001

ISBN 185197 934 6
ISSN 1368 244X

Contents

ACKNOWLEDGEMENTS

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

GLOSSARY OF TERMS

This glossary describes some of the key terms referred to in the paper.

Impact Evaluation
Evaluation to measure the impact a policy or programme has on defined outcome measures. It usually involves measuring the counterfactual (see below).

Process evaluation
A form of programme monitoring designed to determine whether the programme is delivered as intended to the targeted recipients. Also known as *implementation assessment* (Rossi, 1979).

Outcome measures
The outcomes are the factors that the programme is expected to affect. For most active labour market policies typical outcomes would be exit from benefits, entry to work or training, household income etc.

For impact evaluation it is usually crucial to define the outcomes of interest precisely, and in advance of, the evaluation, since this will have a strong bearing on the design of the evaluation. It is also important to reach agreement about the smallest change in the outcomes that would be considered important since this will dictate the sample sizes needed (which may in turn affect the design).

In defining terms such as deadweight and additionality it is usual to talk about 'positive outcomes'. For binary outcome measures, such as entering work, a 'positive outcome' is usually easy to define (although there may be some unclearness about what constitutes work – e.g does unpaid work count and/or work that lasts for no more than a short period?). For continuous outcome measures such as income or 'average time unemployed', a positive outcome would have to be defined in terms of change or relative to a threshold value (e.g. an increase in income, or income over a fixed minimum level).

The eligible population
Those members of the population who are in scope for the programme. For most evaluations the eligible population will be defined in terms of a period of time (e.g. lone parents claiming Income Support between Oct '98 and Dec '98). A distinction might be made between stock and flow.

For voluntary programmes the eligible population is usually defined as those eligible to participate rather than those who do participate. Some evaluations may, however, concentrate on participants, in which case it would be legitimate to refer to participants as the 'eligible population for the evaluation'.

The counterfactual

Also called the *base case*. It is defined as the number of positive outcomes that would have been observed amongst the eligible population if the programme was not in place.

In most evaluations the counterfactual will be measured (with varying degrees of accuracy) using a control group who are not in receipt of the programme.

Deadweight

The numbers or proportion of the eligible population who would have achieved a positive outcome in the absence of the programme. For compulsory programmes this will be the same as the counterfactual; for voluntary programmes deadweight is often defined for participants only. I.e. The numbers or proportion of *participants* who would have achieved a positive outcome in the absence of the programme.

Additionality/programme effect

Number of additional positive outcomes that the programme creates. It equals the number of positive outcomes achieved with the programme minus the counterfactual. It is a measure of the *programme effect* or *impact*.

Displacement/substitution

The change in the number of positive outcomes amongst other non-eligible populations as a result of the programme. The estimation of displacement is, in most instances, extremely difficult, and most evaluations cannot do so with any degree of accuracy. Where an estimate is made it is usually through the analysis of administrative data.

Intervention and control groups

The intervention group (sometimes called the treatment group) is the group in the study who are in scope for the programme. The control group is the group in the study who are excluded from the programme. For some evaluation designs (notably, the matched comparison group design), the intervention group is selected from participants and the control group from non-participants.

Fomative and summative evaluation

Formative evaluation is evaluation undertaken to provide information that will be used to improve a programme; summative evaluation is evaluation used to form a summary judgement about how a programme operated.

1 INTRODUCTION

Evaluations are used in a large number of fields and to answer a very wide range of questions about when and how interventions or treatments 'work'. Although methodologies are to a large extent portable across subject areas, with, for instance, evaluations of health interventions using designs very similar to those used in the evaluation of social policies, this paper concentrates on the particular issues that have arisen in the evaluation of labour market programmes.

The paper provides an *overview* of the evaluation methods currently in use, and should be read as an introduction to the more detailed papers that are to be published in the methodological series. In particular, more detailed papers are to be published on process evaluation and on impact evaluation. The paper is aimed primarily at Government social researchers with a limited knowledge of evaluation methods.

Deciding on a design for an evaluation is a far from easy task. There are numerous approaches that can be used, and the decisions about which evaluation model or models to adopt will depend crucially on the questions of interest and the nature of the policy or programme to be evaluated. At the very basic level, in designing an evaluation the key questions that will need to be considered are:

- Do we need to know more about how the policy or programme operates on the ground? If yes, then what aspects do we need to know more about, and how should information about those aspects be collected? **Process** evaluation, which addresses these questions, is covered in Section 3 of this paper.

- Do we need to know what **impact** the policy or programme has in terms of desired outcomes? And, to answer that question, do we need to know what would happen, should it not be in place? That is, is it necessary to estimate the **counterfactual**? If yes, then how should the counterfactual be measured? The designs described in Section 4 of this paper cover the main methods that might be used.

Sometimes, evaluations may only need focus on either the process or the impact of the policy or programme. More commonly, the research will involve both elements and thus both types of evaluation.

Most of the examples used in this paper are derived from evaluations carried out by the DWP, formerly the DSS, on active labour market programmes. By an active labour market programme we refer to a government policy intervention, usually administered by local Benefits Agency or Employment Service staff, to aid or encourage an entry into or increase in labour market participation.

The paper is divided into three main sections. Section 2 describes the main questions which can be answered by process and impact evaluations, and examines the extent to which the research design is influenced by the parameters of the programme and its timetable. Sections 3 and 4 take us into the detail of the two types of evaluation. Section 3 looks at process evaluation, with a particular focus on why it might be carried out and the possible research methods used. Section 4 is devoted to impact evaluation and, in particular, on methods of estimating the counterfactual. We focus on five methods, namely randomised trials, matched area comparisons, before-after studies, matched comparison group designs and difference-in-differences approaches. Cost-benefit analysis is also discussed.

A key distinction is made in the paper between process and impact evaluation. Another distinction commonly used in the literature, but which we only mention in passing in this paper, is that between formative and summative evaluation. Formative evaluation is evaluation undertaken to provide information that will be used to improve a programme; summative evaluation is evaluation used to form a summary judgement about how a programme operated. The design issues are largely the same whether the evaluation is considered as formative or summative.

2 DECIDING WHICH TYPE OF EVALUATION IS REQUIRED

2.1 When should we use process evaluation?

Listed below are various questions that we may want answered by a programme evaluation. If so, we will want to use some form or forms of process evaluation. Of course, the exact questions that need to be answered will depend on the policy to be evaluated and the objectives that policy makers have for the evaluation. As ever, it is important that the policy objectives are set out clearly so that appropriate research questions can be formulated.

Finding out about service use

- What is the level of awareness among the eligible population and the potentially eligible? How did people hear of the programme? Do they understand the key elements of the programme?
- Do all those eligible receive the programme? Who participates and why; who does not participate and why?
- Do some that are not eligible receive the programme and if so, does this suggest that the target population is poorly defined or that programme delivery is poorly controlled?

Finding out about service delivery

- Is service delivery consistent with the programme's intention?
- Are all elements of the programme being delivered adequately and consistently?
- How much change has occurred since implementation? How does provision compare with that before implementation of the programme?
- Does the programme comply with professional and legal standards? For example, are there appropriate complaints procedures?

Identifying any variations in service delivery

- Are services delivered according to different models or by different organisations (e.g. supplied by BA or ES in some areas and by Contractors in other areas)? If so, how do these compare?
- Are programme resources or programme delivery consistent (and appropriate) across all *geographical* locations?
- Do variations exist across locations or models which provide examples of best practice or which identify important interventions that should be

considered elsewhere? Are there variations over time or for different groups?

- Are services focussed on those who are easier to reach at the expense of those who are harder to help? If so, what is the impact on the nature of the services provided and the net outcomes from these services? What would be the effect of shifting the balance between the easier and harder to help?

Finding out about the organisation of the programme

- Is the programme well organised?
- What are the main models of programme organisation and delivery?
- How well do different groups involved in delivery work together (in terms of different staff within delivery teams, and in terms of different programmes and agencies with which it must interact)?

Looking at the programme's resources

- Are adequate resources being used to deliver the programme?
- Is programme staffing and funding sufficient to ensure appropriate standards?
- Are programme resources used effectively and efficiently?
- Are costs per unit reasonable?
- Are costs per outcome reasonable and are they offset by the benefits?

Looking at participants' experiences of the programme

- Are participants receiving the proper amount, type and quality of services?
- What are participants' experiences of contact with the programme, e.g. how were they invited, what kind/how many contacts did they have, what was the duration/content of their contact?
- Do participants understand the nature of the programme, its intention and its various elements?
- Are participants satisfied with their interactions with staff delivering the programme, with the procedures, and with the services they receive?
- Are there particular groups within the target population which do not receive the programme and why?
- Do participants engage in anticipated or intended follow-up behaviour?

These questions may need to be answered with a variety of approaches including operational research, social research, monitoring, or analysis of administrative data

(e.g. benefit records) and programme databases. Ideally, different sources of evidence should be used to triangulate, verify, substantiate or qualify findings. The main research methods are discussed in Section 3.

2.2 Can process evaluation be used to look at outcomes of the programme?

Although process evaluation is not concerned primarily with outcomes, the methods used (both quantitative and qualitative) often inform learning about outcomes, and may be crucial to the success of an impact evaluation. By and large, information collected as part of a process evaluation will help to address the following sorts of questions, although it may not always be possible to talk in terms of specific quantities (which the impact evaluation is designed to do):

- What effect (positive and negative) is the programme perceived to have on programme participants (and non-participants)?
- What effect does the programme appear to have more widely, for example in terms of cultural change?
- What factors appear to underpin differing impacts and outcomes?
- Are there particular sub-groups within the target group which do better or worse from the programme and why?
- Do particular models of delivery or implementation appear to produce better outcomes? If so, how and why?
- Are there groups for which the programme appears to create more sustainable outcomes?

2.3 When to attempt a measure of the impact of a programme

In Section 4, we outline some of the various ways in which researchers might attempt to measure the impact that a programme has on a relevant outcome or outcomes. In most instances, it is extremely difficult to make an accurate estimate of the programme's impact. To do this, we must measure what would happen - to the relevant outcome or outcomes - if the programme were not in place. This is called, measuring the counterfactual. For programmes where the expected impact is small, the sample sizes needed to carry out impact evaluations are often very large. So, if the counterfactual cannot be estimated using administrative data, then a considerable proportion of the budget for an evaluation will often be devoted to its estimation. It is important therefore, that the need to estimate the counterfactual is carefully assessed before going ahead.

As a starting point, the counterfactual should only be estimated when the primary outcomes for a policy or programme are expressed in terms of change. For instance a policy objective might be written in terms of *increasing* the number of persons in work, or *reducing* the numbers on low incomes. In these instances the counterfactual is the figure from which the increase or reduction is achieved.

In some instances there is a strong case for assuming a particular counterfactual. For example, a new service might be introduced where the main target is to encourage a take up rate above a fixed percentage each year. In this instance the counterfactual would obviously be zero.

A measure of the counterfactual is useful:

- When a good estimate of 'added benefit' or 'additionality' is needed for a cost-benefit analysis
- When there is a need to convince others that a programme or policy is beneficial
- When there are alternative policies or methods of implementation that might be used and there is a need to establish which is preferable.
- When there is a need to establish whether a policy works better amongst some sub-groups than others.

Nevertheless, the counterfactual will be very difficult to estimate reliably if the programme impact is expected to be very small. The problem is particularly acute if the programme impact is smaller than variation that occurs naturally over time in the eligible population.

2.4 Designing an evaluation workable within the context of the programme

Clearly, evaluation research cannot be designed in a vacuum. Evaluations of both the operation and the outcomes of a programme need to be reviewed within context. To a large extent, decisions about the research design will be driven by practical issues such as available resources, the timescale for results and so forth. Arguably the main factors that will determine the evaluation design are the structure and nature of the programme, particularly in terms of: whether the programme is to be introduced nationally or as a pilot; whether participation in the programme is voluntary or compulsory; and whether the programme can be thought of as a single programme or as a complex set of local programmes. In

addition, the emphasis put on different elements of the evaluation will differ depending upon the concerns of those who conceive the evaluation and those who are most interested in the outcomes.

Although these factors will drive the main features of the design there are a large number of other factors that will determine the details. Some of the questions to be addressed are listed below:

- When is the programme to be implemented (immediately or in the future)? If in the future, can and should data for evaluation be collected before implementation?
- When do the findings from the evaluation need to be available?
- What policy changes are anticipated during the evaluation period? How might the changes affect the research?
- Are there different models of delivery that need to be evaluated separately? For instance, are some people offered a telephone help-line while others are given advice face-to-face?
- Are there different sub-groups for which the programme needs to be tested separately? For example, do we need to look across different client groups, service providers, age groups, people receiving different benefits or benefit levels, etc.?
- What geographical areas need to be covered in the evaluation, and what areas need to be avoided (perhaps because another intervention is being piloted there)?
- What are the outcomes that we would like the programme to change? Can these be prioritised?
- What impact is the programme likely to have? What is the smallest impact we would like to be able to detect? What is the smallest impact that will be deemed to be a success (or failure)?
- What are the factors that will affect the impact?
- What are the participation rates likely to be?
- What data is currently available that the evaluation can make use of? What extra will need to be collected? How will that data be collected?
- Are monitoring processes established and collecting satisfactory data?
- What resources are available? This applies to different types of resources, including money, staff, time, central Government staff, etc.

The answers to these questions will largely determine how the evaluation is designed. Nevertheless, the various requirements of the research may mean that

the design process goes through several stages before a satisfactory design is found which will address all the necessary questions of the evaluation.

2.5 Persons to be included in the evaluation

An important decision that needs to be made during the early stages of the development of an evaluation is who are the 'subjects' of the evaluation. That is to say, which actors in the process in question need to be studied in order to effectively answer the questions set out in the evaluation?

The relevant actors can be considered in three main groups:

(i) Primary subjects

These are the individuals to whom the policy or programme is specifically targeted or who are directly affected by the policy or programme. Examples are -

- Eligible / target population
- Claimants
- Recipients
- Non-recipients (perhaps divided into those who have or have not claimed the benefit, eligible non-recipients, potentially eligible non-recipients etc.)
- Participants
- Non-participants (who have or have not engaged with the programme)
- Leavers (from a benefit or programme)

(ii) Secondary subjects

These are the individuals who may have a key role in making the policy or programme work but are not integrally involved in the development or delivery of the policy or programme. For example, in the New Deal programmes, employers are crucial because they supply the work opportunities that make the New Deal effective. Representatives of these groups may be involved to a lesser or greater degree in the strategic development of the policy or programme and these too may be incorporated in the evaluation. Examples are -

- Employers
- General Practitioners
- Pension managers
- Intermediaries such as CAB, advice centres, referral agencies
- Other subjects indirectly affected by a policy that is not delivered by them, for example a hospital emergency wards affected by people being encouraged to visit their GP.

(iii) Actors involved in delivery

One of the key groups usually studied in evaluations are those involved in the delivery of new policies or programmes. Furthermore, the new emphasis on cross-departmental working and partnerships, means that additional actors and agencies are increasingly involved in the delivery of policies and programmes and so are included in evaluation designs. Examples are -

- Policy makers / implementers
- Staff involved directly in delivery e.g. programme managers, personal advisers, adjudicators, independent organisations contracted to deliver services
- Secondary actors involved in delivery e.g. Employment Service staff, Local Authority staff and independent contractors.

3 PROCESS EVALUATION

3.1 Uses of a process evaluation

Almost all large scale evaluations of government policy will include some elements of process evaluation since, arguably, this is the side of the evaluation that provides most information on how the policy should be managed or developed in the future. In formal terms process evaluation (Scheirer, 1994) "verifies what the program is and whether or not it is delivered as intended to the targeted recipients".

Process evaluation tends to be carried out independently of service delivery and is normally carried out by evaluation specialists. For practical reasons, it is normally separate from day to day programme management and monitoring, though data from Management Information Systems can play a vital role and be built into the design of the programme to allow for evaluation.

In some instances a process evaluation may be all that is needed for an evaluation. Possible scenarios where this might arise are:

- for a relatively new programme where further development is likely before an assessment of the impact of the final model is needed. In this instance the process evaluation would be considered as 'formative';
- with an established programme which is under-performing or where questions have arisen about organisation, delivery, quality, or success;
- with a programme where effectiveness is known or assumed and only the implementation and delivery is in question;
- where impact evaluation is ideally wanted but the size of the programme, its expected effect, or the length of time available to allow outcomes to emerge are too small to make outcome evaluation possible.

More typically a process evaluation will be carried out alongside an impact evaluation. In these instances the process evaluation data can be used independently of the impact evaluation to assess the implementation procedures. But perhaps just as importantly, the process evaluation provides very useful contextual data against which the results from the impact evaluation will be judged. It is common for instance, for an impact evaluation to give one overall figure for the impact of the programme on the eligible population in aggregate, but for the sample numbers to be too small to allow for reliable estimates on impact on

sub-groups to be made. In these instances the process evaluation findings can often be used to make judgements (sometimes qualitatively rather than quantitatively) about where and amongst who the impact of the programme is greatest.

3.2 Methods of process evaluation

Early sections of this paper have discussed the role of process evaluation and the answers which it may address (Section 2.1). The issue about who should be included within the evaluation has also been discussed (Section 2.5). In this section, we look at the variety of **research methods** available for answering the questions and communicating with the relevant actors. It is important to establish what method or methods of data collection will be most effective within each evaluation. Each method will of course carry with it particular advantages and constraints (as well as cost implications). Often, a combination of methods is necessary to answer all the questions posed by an evaluation. Evaluations which combine methods successfully are also likely to be more robust, where they are able to combine findings pertaining to the same topic from different sources. They will also, of course, often be the evaluations that are the most complex to analyse and interpret.

The range of methods might be divided into two broad groups, namely monitoring / operational research and social research, the key distinction for our purposes here being that monitoring and operational research use data collected for purposes other than the evaluation proper, whereas social research methods are usually used to collect data specifically for the evaluation. The two methods are described in turn below.

3.2.1 Monitoring and Operational Research

Monitoring and Operational Research may involve:

- The analysis of administrative data about the eligible population e.g. benefit records
- The analysis of data from Management Information Systems e.g. programme database
- The collection and analysis of performance measurement data about resources (staff and financial), caseloads, etc.
- Special monitoring exercises e.g. *pro formas* attached to specific cases to identify stages and timing
- In-depth area studies

It is usual for Government departments to use administrative data to monitor the progress and, where possible, to estimate the impact of programmes. In addition Operational Researchers will often carry out specific modelling tasks, examples of which include:

- Analysis of customers' behaviour, in terms of their interactions with the organisation, during and subsequent to their involvement in the programme. This involves building models of the process of interaction.
- Statistical analysis of data collected from IT systems, or from customer surveys, to establish optimal methods for encouraging participation.

3.2.2 Social research

Social research, used in the context of evaluation, might involve:

- Large-scale quantitative surveys
- Qualitative approaches using depth and paired depth interviews or discussion groups
- Case studies
- Literature reviews, observational studies, diaries and documentary analysis.

Employing social research methods inevitably adds to the costs of an evaluation since they involve primary data collection. They are used when it is considered that monitoring and administrative data is either incomplete (for instance, if it provided no information about employers) or insufficient.

Below we describe the three main methods of data collection of social research: quantitative surveys; depth interviews and group discussions; and case studies.

3.3 Quantitative surveys

The main role of surveys in evaluation is to collect quantitative data on (usually large numbers of) the subjects of the evaluation (see Section 2.5), such as participants and local employers. Surveys of those involved in the delivery of the programme are much less common because of the small numbers involved.

Any surveys, especially surveys of the eligible population, will usually be designed to answer both process and impact questions, and this dual role will largely determine survey design and sample size. Some of the conflicts that are likely to arise are described below:

- For process evaluation it may be important for surveys to be 'representative' of the eligible populations in the areas the programme covers. Impact evaluations are more likely to need samples that are comparable with control groups than are representative per se.
- For voluntary programmes participants are likely to be of more interest for a process evaluation than are non-participants (although some process questions will be specifically targeted at non-participants). There may therefore be a strong argument for over-sampling participants in any survey. Depending upon the design of the impact evaluation this over-sampling may or may not be appropriate for the impact estimation.
- Similarly, for a process evaluation there is likely to be some value in stratifying the survey sample in a way that maximises the possibility for sub-group analysis (either in terms of the characteristics of the eligible population, or in terms of the type of programme intervention delivered). This may conflict with the needs of the impact evaluation.
- Pre-participation surveys may, for process evaluation purposes, need to cover questions about awareness and attitudes towards the programme. Focussing attention on the programme may however contaminate the impact evaluation.

The needs of the process evaluation are likely to drive factors such as interview length and mode. For instance, it is likely that a short telephone interview would be sufficient in many cases to collect information on the outcomes needed for the impact evaluation. In contrast, the demands of a process evaluation are more likely to point to a lengthy face-to-face interview. Furthermore, if an aim of the process evaluation is to examine how the programme changes over time then this may require several surveys rather than just one. In contrast, the need for very large sample sizes is likely to be driven by the demands of the impact evaluation.

3.4 Qualitative research: depth interviews and group discussions

Qualitative research (in particular depth interviews and group discussions) provides in depth investigation of the nature of social and organisational behaviours and how and why they occur. It is characterised by the use of exploratory and interactive methods of data collection that aim to capture the form, complexity or origins of the issues being reviewed.

Qualitative approaches are considered particularly appropriate for process evaluation because they enable researchers to explore, in great detail, the efficacy of the organisation and delivery of a programme. These approaches allow

evaluators to assess the features which are appraised as more or less effective by those directly involved in either the design, delivery or receipt of the interventions. In addition, qualitative approaches allow for an investigation of the range of factors which can affect overall outcomes and provides detailed exploration of the factors underpinning participants' experiences of programmes.

Qualitative research is very different to survey research and has very different uses in evaluation. Whereas surveys provide the hard quantitative estimates needed to answer the 'how many' questions, but fail on many of the depth questions about 'why' and 'how', qualitative research cannot give quantitative estimates, but does allow many of these more depth questions to be addressed. Because of these different focuses, qualitative research is often used alongside surveys. Furthermore, because of the small number of personnel involved, qualitative research is the main tool used amongst programme delivery staff. Most evaluations will, however, also include qualitative research amongst participants even when participants are covered by surveys.

Qualitative process evaluations can involve a range of different methods of data collection. Popular tools include individual depth interviews and group discussions. Other approaches which might be employed include paired depth interviews, observational work and documentary analysis.

Individual or paired depth interviews use a topic guide (or interview schedule) which lists the key themes and sub-themes to be explored. Individual interviews are an ideal forum to explore detailed personal experiences of interventions allowing respondents to describe and evaluate their personal experiences of the intervention. They are also suited to eliciting personal or sensitive information.

Group discussions (also called focus groups) usually involve around six to eight respondents, with a moderator or two co-moderators. Group discussions provide an appropriate area to bring together participants/recipients or service providers to discuss, share and compare their experiences of the intervention. The exchanges that occur between respondents can highlight common experiences and views, identify differences within the group, and act as a stimulus to further thought among respondents. They are also a stimulating environment for generating solutions and strategies. However, they are less suitable where the evaluation requires exploration of detailed personal accounts such as employment histories or detailed experiences of service delivery.

It is worth noting the special role of *longitudinal* qualitative research in process evaluation. Longitudinal elements can help to identify changes to the delivery, practices, and organisational management of interventions / services / programmes. In particular revisiting a previous sample can help to identify the nature of changes over time to attitudes, decision-making , behaviours; the factors influencing the durability of outcomes; and provide reflective perspectives on the experiences of different sub-groups within the sample. A paper on the use of longitudinal qualitative research in evaluation is to be published in the series.

3.4.1 Examples of the use of qualitative research

New Deal for Young People (NDYP)

The evaluation of NDYP included a series of six qualitative studies amongst participants to review different stages of the programme. The research objectives were to: to provide information about participants' experiences of the different components of New Deal and their appraisal of them ; to examine the nature of the impact of the programme on participants' aspirations and employment or educational outcomes ; to identify reasons for leaving the programme and / or the unemployment register and identify the factors associated with different routes of ; and, to identify the factors that facilitate or might enhance the effective organisation and delivery of New Deal. Because both in-depth personal accounts and comparisons between participants were required, the study included both depth interviews and group discussions.

The Educational Maintenance Allowance (EMA)

EMA (a DfES programme) provides financial incentives to school leavers to encourage participation in post-compulsory education. A series of extension pilots are being implemented which aim to provide additional, or different forms of, financial support for students with children, disabled students and students who are homeless.

As part of the evaluation programme a quantitative survey was carried to establish the factors which influence the choices young people make following their post-compulsory education (PCE) at 16. A qualitative study followed-up the survey respondents exploring in greater depth their decision-making at this time, views, experiences and impact of participation in EMA. The key aims of the qualitative study were to: explore attitudes to post-compulsory education ; identify factors which influence participation in post-compulsory education; explore attitudes to and experiences of the EMA scheme; investigate the impact of EMA on participation, retention and achievement in post-compulsory education ; explore the impact of EMA on financial decision-making, expenditure and transfers within households; compare and contrast different EMA variants; *and to* identify suggested changes and improvements to EMA. A qualitative approach was chosen to provide a detailed understanding of the operation of EMA scheme. The study was carried out using depth interviews amongst a purposively selected sample of young people and their parents. The design also incorporated a longitudinal element to allow the longer term impacts and outcomes from the scheme to be explored as a complement to the Year One process evaluation data.

3.5 Case studies

The focus of depth interviews and group discussions is the experience of *individuals* seen from their perspective. Case studies, in contrast, look at individuals or organisations from the multiple perspectives of key actors. These perspectives help to build a detailed understanding of the experiences and outcomes in a specific case, where a case can be an individual client / participant; an area; an office etc.

Case studies are generally included in an evaluation when the evaluation requires a detailed understanding of these multiple perspectives on an intervention. For example a case study might be used to explore, in detail, an individual client's experiences of a programme where the roles or accounts of different actors are seen to affect the delivery or impact of the programme on that individual. Alternatively, case studies may be appropriate when exploring factors accounting for different models of delivery where multiple actors are involved in designing or implementing an intervention.

3.5.1 Examples of the use of case studies

New Deal for Disabled People

In the evaluation of the New Deal for Disabled People pilot, case studies were used to examine the ways in which partnerships for delivery were operating in different pilot areas. Each case study involved depth interviews with key actors from different organisations within the partnership.

Education Maintenance Allowance

EMA (a DfES programme) provides financial incentives to school leavers to encourage participation in post-compulsory education. A series of extension pilots are being implemented which aim to provide additional, or different forms of, financial support for students with children, disabled students and students who are homeless.

The evaluation of EMA (see Section 3.4.1) involves case studies of individuals. Each case study involves the investigation of the circumstances and key issues particular to a single individual students' case by seeking the multiple perspectives of key actors. Each study involves a depth interview with the student and depth interviews with nominated key support workers, family members or significant others.

4 IMPACT EVALUATIONS, OR METHODS OF ESTIMATING THE COUNTERFACTUAL

As we have outlined earlier, the primary aim of an impact evaluation is to measure whether a particular programme has achieved its desired outcomes. To do this, we measure outcomes with the programme in place and compare then to outcomes without the programme (i.e. the **counterfactual**). Measuring outcomes with the programme is relatively straightforward since it is largely a matter of observing what happens. Measuring the counterfactual is much more difficult since no direct observations can be made.

A variety of different methods have been proposed for estimating the counterfactual. The feature all the methods have in common is that they compare a groups of people involved with the programme – which we term the 'intervention' or 'treatment' group – with a group of people who are not involved - the 'control' group. The counterfactual is then estimated as the outcomes for this control group. The intervention group will usually be everyone who is eligible to be involved in the programme (the 'eligible population'), or a sub-sample of the eligible population, for instance people who are actually participating in the programme. In principle, the control group should be people who are, on average, identical to the people in the intervention group, with the single exception that they are not involved in the programme. Much of the discussion about the relative merits of different evaluation methods is centred around the extent to which the control group can be said to truly mirror the intervention group in all areas apart from the programme intervention.

The only design where this assumption of equivalence between the intervention and control groups is guaranteed is the 'randomised trial' (often referred to as an 'experimental design'). This is the design considered first (sub-section 1) in the pages that follow. All other designs are referred to as 'quasi-experimental'. The ones we consider here are the matched area comparison design (sub-section 2), the before-after study (sub-section 3), the matched comparison group design (sub-section 4), and difference-in-differences (sub-section 5).

A short discussion of the role of cost-benefit analysis in impact evaluation is included as section 4.6.

Sections 4.1 to 4.5 below focus on the theory and logistics of each approach. For most designs the basic data collection procedure is the same: the two groups (intervention and control) are identified at one point in time and they are then

followed up over time. After a suitable interval outcomes are measured, using either survey or administrative data. The outcomes for the control group give the estimate of the counterfactual; the difference in outcomes between the control and intervention groups gives an estimate of the programme impact.

Before turning to the main impact evaluation methods it is worth mentioning one relatively crude method that can be used in conjunction with other methods, which is to ask participants in a programme to say what the counterfactual would be. If participants can assess what would have happened to them if they had not participated, then an estimate of the counterfactual can be derived directly from those assessments. As an example of this, on the NDLP prototype, participants who found work were asked to say whether or not they would have found that work without the help of the personal advisor. Those saying 'yes' provide the counterfactual. It is unlikely that in this instance the estimate of the counterfactual would be given much credence (unless supported by other evidence) since it is probable that the assessment is more a statement about how helpful the participant found the personal advisor rather than a genuine assessment of the help given. However, there may be other advice given by the personal advisor for which the participant would be more able to assess the impact of the personal advisor. For example, did the personal advisor put you in contact with training providers who you were not previously aware of? Or, did the personal advisor give advice on how to change your job search activities that you followed? A very useful role for this general approach would be to establish, once the *overall* programme impact has been established, whether or not a programme appeared to work better for some sub-groups than others.

It should be noted that the main objectives of each design discussed below is to estimate the direct impact of the programme on the population eligible for the programme. It is, however, perfectly reasonable to expect that the introduction of labour market programmes that focus on particular populations will have an impact, often negative, but sometimes positive, on other populations as well. Evaluation designs rarely include the ability to measure these displacement or substitution effects directly, primarily because the effects will usually be thinly dispersed which makes detection of the effects very difficult. In designing an evaluation some consideration should be given to whether or not this omission is important: if it is then the evaluation may have to be broadened. The implications of this on the evaluation design are beyond the scope of this paper, but are to be covered in the forthcoming paper in the series on impact evaluation.

The discussion that follows on each design is necessarily brief. A more detailed consideration of the various methods is also included in the forthcoming paper.

4.1 Randomised trials – an experimental design

4.1.1 The main features of a randomised trial

In a randomised trial the eligible population are assigned *at random* either to an intervention group or to a control group. The control group are denied the programme and are treated, as far as is possible, as if the programme did not exist. The control group usually receive existing services.

There are two main types of randomised trial: randomisation of individuals within areas; and randomisation of areas. In the latter whole areas are either assigned to the intervention or control group.

4.1.2 The designs' strengths and weaknesses

The randomised trial is considered the 'gold standard' for evaluation. However, it is not always used in practice. This is because of practical difficulties in implementation rather than because some other design is more powerful or reliable.

The main strength of the randomised trial or 'experiment', compared to quasi-experimental designs, is that the only two differences between the intervention and control group are 'random differences' and 'the influence of the programme'. Systematic differences, such as differences in motivation between the members of the intervention and control groups, can be ruled out.

A further advantage is that there is no necessity to do a 'before-programme' study to check that the intervention and control groups are alike. (See Section 4.5 on difference-in-differences.)

Countering these advantages, are the difficulties (and sometimes impossibilities) in allocating people randomly within a particular area to an intervention or control group. Firstly, it often means running two administrative systems within a single area. Secondly, for voluntary programmes, there is an ethical difficulty in advertising programmes that some people will be denied entry to if they apply. In some cases, such as whole community interventions, such as a Health Action Zone,

it is **impossible** to randomly allocate people in the same area to an intervention or control group.

Randomly allocating areas, rather than people within areas, to intervention or control groups is typically ruled out because of the large number of areas that would be needed. The minimum sample size of areas would be about 20 in the intervention group and 20 in the control group although in some instances the numbers needed may be much greater. Also, compared to a 'within-area randomisation' the sample size of people within the study would also have to be significantly larger because if the need to allow for between-area sampling variation. This has cost implications if the data on outcomes are to be collected via surveys.

Although the results from randomised trials have a credibility that other evaluation designs do not, the estimate of the programme effect may be biased if any of the following occur:

- if some of those control group do actually participate in the programme;
- if surveys to measure outcomes get differential non-response in one or other of the groups;
- if being denied access to a programme makes those in the control group act differently to how they would act in the total absence of the programme;
- if those randomised to the programme affect the behaviour or attitudes of those in the control group (which might happen if members of the two groups know one another);or
- if the trial affects the way the programme is implemented.

If the effect of the programme is expected to be small and/or participation rates are low, a randomised trial of the eligible population will need a large sample size. One possible exception to this rule is if allocation to the intervention and control groups can be done at the point of participation.

4.1.3 Practical issues

Most of the practical issues that arise in randomised trial designs occur only if people are divided into intervention and control groups within their area, rather than if whole areas are divided into intervention or control groups. The main issues are:

- Very good management systems need to be put into place to ensure that local administrators know who is in the intervention and who is in the control

group. Systems are needed to monitor, and help avoid, programme participation by any of the control group.

- It is not always immediately apparent who is eligible to be included in the randomisation, and who is not. For voluntary programmes, we could decide to include the whole target population, or only those expressing an intention to participate, or those actively trying to participate (or, in principle, at any other point along the continuum between 'eligibility' and 'active participation'). The closer we get to 'active participation', the smaller the trial can be. However, the costs, practical and ethical difficulties of identifying those reaching a particular point on the line may offset any such advantage. Furthermore, the impact of the programme on those excluded from the randomisation cannot be known. (So if randomisation is done at the point of participation, the impact of the programme on non-participants will not be measured.)
- The size of the control group within each area has to be kept relatively small if the programme being offered to the intervention group is to be a reasonable approximation to the programme that would be offered to everybody. This will tend to increase the number of areas that the trial needs to be run in.
- It may be considered unethical to extend a trial beyond the point at which a positive benefit of the programme has been demonstrated. But because at the end of a trial the control group will usually be free to participate, the trial then gives no evidence on the long-term impacts of programmes. Longer-term benefits tend to be estimated using quasi-experimental methods.

4.1.4 Examples of randomised trials

Intensive Gateway Trailblazers (IGTs)

For all IGTs clients were randomly allocated to either an IGT group or to a control group. The randomisation was done using NINO digits: clients with an odd NINO were assigned to the IGT group; those with an even NINO were assigned to the control group.

Fully Fledged Employment Zones

In the impact evaluation of the Fully Fledged Employment Zones two methods were used to estimate the counterfactual. Within four of the zones individuals who were eligible for the zone were randomised to one of two groups: intervention and control. As with the AGTs the randomisation was done using NINOs. The remaining 11 zones were evaluated using a matched area comparison design (see Section 4.2). In the 11 zones all those eligible for the zone were included in the programme, and these 11 areas were matched to comparison areas with similar labour market characteristics. For both methods data on outcomes was derived from administrative sources, namely MI, JUVOS, and LMS.

NDLP Prototype

The impact evaluation of the NDLP prototype was designed as a matched area comparison (see Section 4.2) but within each area the invitations to join the programme were sent out over a period of time and in essentially random order, the date the invitation was sent being determined by NINOs. For the early part of the prototype this mimics a randomised trial, those invited early being the intervention group and those invited later being the control group. In practice some of the randomisation was lost because members of the 'control' group were allowed to self-refer onto the programme. Nevertheless, the results from this 'natural' experiment were considered by some to provide a more robust estimate of the counterfactual than the matched area comparison.

4.2 Matched area comparison design

4.2.1 The main features of a matched area comparison design

In this design, a new programme or policy is piloted within a **small** number of selected pilot areas (typically about 10). These areas are then matched to a set of 'control' areas, usually, although not always, on a one-to-one basis. The control areas are carefully selected to match the pilot areas in terms of their labour market characteristics.

The population eligible for the programme is identified within each selected area – both pilot and control. Either all the eligible population or just a sample is followed up. Then, after a suitable interval, outcomes are recorded. Differences in outcomes between the pilot areas and the control areas are attributed to the programme. Typically this is after controlling for any major differences between the eligible populations of the areas.

4.2.2 The designs' strengths and weaknesses

The primary strength of the matched area comparison is the ease with which it can be implemented. It avoids most of the problems of randomly allocating people to intervention and control within areas. Only one administrative system per area is needed, and no individuals within pilot areas are denied access to the programme. For this reason, matched area comparisons are usually considered more ethically acceptable than randomised trials of individuals within areas.

The main weakness of the design is that the results from it are often very difficult to interpret. The difficulty arises because of the main feature of the design, namely that the control group members all live in different areas to the pilot group

members. This means that any observed difference between the pilot and control groups might, potentially, be attributable to three things -

- Differences in the labour markets in the pilot and control areas;
- Differences in the eligible populations of the areas (these might include differences in factors such as: motivation to find work; ability to look for work, and qualifications for work).
- The programme itself.

To estimate the impact of the programme itself, we have to control for both of the first two factors.

In principle differences in labour markets are controlled for by carefully matching the pilot and control areas at the selection stage. But, no matter how carefully this matching is done, residual differences are likely to remain. For programmes targeting special populations who experience the labour market in non-standard ways there is the additional difficulty of being uncertain about what aspects of the labour market it is important to match on. Furthermore, there is the problem that a good match at the start of the programme may prove to be a much poorer match at the time outcomes are measured. Changes in specific areas may occur over time that make them different at the end of the study. For these reasons 'mis-matching' has usually to be controlled for at the analysis stage. With a small number of areas this can only be done crudely.

Differences between the people eligible in the pilot and control areas, such as personal circumstances (eg. child care responsibilities), previous work experience, educational qualifications, and motivation to find work, are usually controlled for at the analysis stage using statistical modelling. If the analysis is based on administrative data, then the ability to control for these factors is very limited. There is more potential with survey data, but the estimates of the programme effect may be sensitive to the adequacy of the models used and the reliability of the data.

The problem of controlling for these factors is particularly acute when the programme effect is expected to be small. As an example, the NDLP prototype estimated that the number of lone parents finding work was about two percentage points higher in the prototype areas than in the control areas. It is difficult to feel confident that this difference is entirely attributable to the programme and not, at least in part, attributable to labour market and/or population differences between the prototype and control areas, for which the statistical modelling could only partially control.

4.2.3 Appropriate uses of a matched area comparison design

This design is appropriate when the expected impact of the programme is considerably greater than any variation which we might expect between the areas if there was no programme in place (and after controlling for labour market and personal characteristics). This variation will rarely be known exactly, but it will often be possible to make an estimate using a few years of historical administrative data (with 'leaving benefits' as one possible outcome measure).

In estimating the expected impact of the programme – the 'expected programme effect' - account needs to be taken of the programme participation rate. A high impact on participants, but a low participation rate, will lead to a low overall programme impact. For matched area comparison designs, which compare *eligible* populations between areas rather than compare participants, low participation rates can make an impact almost impossible to detect. (Other designs, such as the matched comparison group design (see Section 4.4), measure the effect on participants only and so low participation rates are less problematic.)

4.2.4 Improvements on the basic matched area comparison design

The basic design might, in some circumstances, be improved by:

- Increasing the number of areas in the study. This might be achieved by implementing a pilot programme within areas smaller than BA districts, but using more BA districts.
- Incorporating a 'before-programme' estimate within both the pilot and the control areas. This might be achieved using administrative data or by collecting retrospective work histories from sampled members of the eligible population. (Essentially this approach means extending the design to incorporate a difference-in-differences estimate – see sub-section 4.5).

4.2.5 Examples of a matched area comparison design

NDLP Prototype

For the evaluation of the NDLP Prototype the eight prototype areas were matched to six 'control' areas. Outcomes were measured in terms of exits from benefits (which could be monitored using administrative data, and on the whole of the eligible population), and in terms of entry to work (which was monitored using survey data on sub-samples of the eligible population). In the event no significant differences were found between the prototype and control areas and it is likely that this was because of the relatively low participation rate (at about 20%): an estimated 10% additional employment amongst *participants* equated to just a 2% difference in entries to employment between the eligible populations of the prototype and control areas. Since the matched area comparisons findings were largely consistent with those of the NINO comparison (see the example in Section 4.1) it was possible to attribute this 2% difference to the programme; without the NINO comparison the 2% might just have easily have been attributed to a residual difference between the prototype and control areas.

Earnings Top-Up (ETU)

In the evaluation of the Earnings Top-up, the benefit was introduced in eight pilot areas. These eight areas were matched with four control areas, which were very similar in their characteristics with the exception of the benefit introduction. Within the twelve study areas, surveys were conducted of employers, low paid workers in work and the medium-term unemployed. Outcomes were measured in terms of net income and entry into work. The evaluation also incorporated a longitudinal element, resulting in a differences-in-differences design, discussed below in Section 4.5.

ONE

The ONE evaluation also uses the matched area comparison approach. ONE has been introduced into twelve pilot areas, and these twelve areas have been matched to control areas (twelve in total, but not using a one-to-one match), where ONE has not been implemented. The outcomes of the participants in the pilot areas will be compared with those for similar people from other, comparable areas.

4.3 Before-after study

4.3.1 The main features of a before-after study

Before and after measurements can be incorporated into most evaluation designs. What we refer to here as a 'before-after study' is just one particular instance of their use.

In a standard before-after study, outcomes will be measured on the population eligible for a programme both **before** the programme is implemented and **after**. The difference between the before and after measurements is taken to be the impact of the policy. (In this instance, the 'before' – or 'baseline' - measurements act as the control measurements.)

Typically outcomes are measured at just one point in time before programme implementation and at one point in time after implementation. But this basic design is considerably strengthened if the number of measurement occasions is increased both before and after.

Before-after studies are primarily used in instances where a policy is to be implemented nationally without a pilot stage.

4.3.2 The strengths and weaknesses of a before-after study

The key strength of the before-after study is that it is possible, in theory, to implement a policy nationally and yet still obtain a measure of the impact that policy has. In practice, the design has considerable weaknesses that make its use very problematic.

The main weakness of the design is that change brought about by the policy cannot be separated out from change that would have happened anyway (i.e. 'natural change' or change brought about the introduction of other policies at about the same time). This is particularly problematic if the policy is expected to have a relatively small impact, the worst case being where the expected change due to the policy is smaller than the change that happens 'naturally' from year to year.

The design can be strengthened quite considerably if the time series is extended to several years (or periods) before the implementation of the policy and several years after the policy is implemented. It then becomes possible to look for an 'interruption' or 'shift' in the time series at the time the policy is introduced and to check that the shift is sustained over time. However, although this is a relatively powerful approach, the strong data requirements means that it is usually only possible to use administrative data or other standard datasets (such as large repeated government surveys). This limits the outcome variables that can be used. Furthermore, the approach means that results on the success of the policy will only be available some considerable time after the programme is implemented.

4.3.3 Design issues

The before and after measurements can be taken on different cross-sections of the people eligible or by taking repeated measurements on the same people within the eligible population. One variation on this is to select a sample once a policy has been implemented, and then to collect data on that sample retrospectively.

Although in instances where the eligible population is reasonably constant over time, it is often preferable to take repeat measures from the same sample - in that it makes measures of change more precise - there is a danger that change over time will include not only 'natural change' and change due to the programme, but also include a 'age-effect' (i.e. an effect due to the fact that that the sample members are older 'after' than they were 'before') . Repeat measures data should consequently be used with extreme caution.

4.3.4 Appropriate use of the before-after designs

Before-after and time series approaches are probably of most use as a supplement to other more formal evaluation methods. In particular, the administrative data collected by government departments allows for long-term time series analyses to be carried out relatively easily and cheaply. The results from these analyses play a very useful role in validating and confirming the conclusions from the short-term intensive evaluations carried out by independent researchers.

4.3.5 Examples of before-after designs

Jobseekers' Allowance (JSA)

To evaluate the impact of 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.

4.4 Matched comparison group design

4.4.1 Main features of a matched comparison group design?

In a matched comparison group design, the people eligible for a programme (the eligible population) are divided into two self-selecting groups - participants and non-participants of the programme. An intervention group is then selected from the population of participants and a control group is selected from the population of non-participants. The controls are selected so that they 'match' the participants.

The usual approach is to match each selected participant uniquely to a non-participant. If the matching is done well, the non-participant controls will be identical to their matched participants in all relevant respects with the single exception that the participants participate and the controls do not.

'Relevant respects' for matching are factors that are associated with participation and with the outcome(s) of interest. For example if the outcome of interest is 'finding work' and participants are disproportionately self-motivated well-qualified people whose personal circumstances make working a feasible option, then the non-participant controls should have the same characteristics.

Once matches are found, both participants and their matches are followed-up over time and outcomes recorded. The difference in outcomes between the participant and control groups gives an estimate of the impact or effect of the programme.

4.4.2 The strengths and weaknesses of a matched comparison group design

The matched comparison design has a number of strengths:

- It is a very useful evaluation method in the instance where a programme has been introduced nationally (if done well it should be less biased than a simple before-after study).
- There are few, if any, ethical difficulties because nobody is denied the programme.
- Unlike the matched area comparison where the intervention group is everyone who is eligible for the programme, for whom overall programme effects are likely to be small, the matched comparison group design concentrates on participants (where arguably all of the programme impact is concentrated). As a consequence the sample sizes needed for a matched comparison design are often considerably smaller than those needed for a matched area comparison. If outcomes are to be collected using survey data this can significantly reduce the costs of evaluation.

On the down side, the design is very dependent upon the matching procedure. Any inadequacies in this can introduce bias into the estimates of programme effects. As an illustration of the problem, let us suppose that NDDP participants are matched to non-participants on age, sex, disability, and duration on benefits. But suppose that, within an age, sex, disability, and duration combination, participants are more likely to be actively seeking work than are the non-participants. If this is the case, even without the programme, the participants will be more likely to find work than their matched non-participants, and this is the case irrespective of their participation. In this scenario the effect of participation (i.e. the programme effect) is likely to be over-exaggerated.

4.4.3 Practical issues

To do the matching well means that all of the relevant matching variables have to be identified, plus reliable data needs to be available on all of the relevant variables for at least a sample of participants and non-participants.

Furthermore, if some of the relevant variables are likely to change with participation (one example might be 'confidence about finding work'), and the information on these variables cannot be collected retrospectively, then the information will need to be collected prior to participation. For programmes with low participation rates this may involve a very large data collection exercise amongst the eligible population.

The actual procedures for matching are fairly technical, a number of options being available. The method that is most likely to be used at present is 'propensity score matching'. Propensity score matching will be the subject of a separate paper in the methodological series but, in broad terms, the method works by fitting a logistic regression model of the probability of participation using, as predictors, factors thought to impact on both participation and outcomes. Participants are then matched to non-participants on the basis of this modelled probability or propensity score. The method is an attempt to mimic randomisation at the point of participation, the principle being that the participants and their matched non-participants are just as likely as one another to participate. In each matched pair the decision on which of the two participates is treated as essentially a random process.

A matched comparison group design is only likely to be feasible for programmes with fairly low take-up rates since only then will there be an adequate pool of non-participants from which adequate matches can be found.

4.4.4 Examples of the matched comparison group design

NDLP National

The evaluation of the NDLP National is using a fairly sophisticated version of the matched comparison design. A very large (65,000) sample of eligible lone parents who have not participated in the programme were sent a postal questionnaire which collected information on their characteristics, job experience, qualifications and attitudes towards work. Members of the sample who participated within a short fixed period formed the 'intervention group'; each of these was matched to another member of the sample who did not participate, the matching being done using a propensity score modelled using the responses to the postal questionnaire. After a suitable period of time the participants and their matched controls are to be interviewed to establish whether or not they have entered work.

Training for Work (TfW)

A simpler version of the NDLP National design was used to evaluate TfW. In this instance a random sample of people leaving TfW were matched to non-participants, the matching being based on variables available from administrative systems. The main distinction between the NDLP National and TfW evaluations is in the much richer material from which matches can be made in the former.

4.5 Difference-in-differences

4.5.1 The main features of the difference-in-differences design?

The difference-in-differences approach is more appropriately considered as a method of estimation rather than as a design alternative in its own right. The approach can be combined with all of the preceding designs discussed, with the exception of the straightforward before-after design.

In a difference-in-differences approach, two groups are compared both before and after a programme or policy is implemented. Typically the two groups will be participants and non-participants from the same eligible population, but the two groups could be the eligible population and some other population (which might be the eligible population from a control area). In all cases, one group represents the 'intervention group' and the second is the 'control group'.

The idea behind the approach is that two measures of change over time (i.e. 'differences') are calculated - one for the intervention group and one for the control group. The difference for the control group gives an estimate of the change over time that would have happened if the programme had not been introduced (i.e. it measures 'natural' change). The difference for the intervention group is a measure of this 'natural' change plus change due to the introduction of the programme.

Subtracting the difference for the control group from the difference for the intervention group gives an estimate of the change due to the introduction of the programme or policy.

There is an implicit assumption that the size of the 'natural change' is the same for the control and intervention groups.

4.5.2 Illustration of the difference-in-differences approach

Consider a voluntary programme where the main objective is to help people into work. One potential measure of success would be a higher than average time spent in work with the programme than without the programme.

A difference-in-differences approach might be to divide the eligible population into two groups: participants and non-participants. A sample would be taken from both of these groups (possibly with matching on administrative data variables such as age, sex, length of time of benefits etc.).

Data on work histories would then be collected from both samples and for two periods: a fixed period before the start of the programme, and a fixed period after the start of the programme. (This data might be collected using administrative records, if available, or by survey.)

For all sample members two figures would be estimated: e.g. percentage of time spent in work in the period before the start of the programme, and the percentage of time spent in work after the start of the programme. The difference between these two figures would then be calculated giving an estimate of change per person.

The average of these individual differences would be calculated per group. The difference between these two averages would be the estimate of the programme impact.

In practice the analysis would be more sophisticated than this, with, perhaps, the differences being modelled. But the principle remains the same.

4.5.3 The strengths and weaknesses relative to other designs

The difference-in-differences approach will, in most instances, give a better estimate of a programme effect (i.e. additionality) than the simple **before-after design**, since it makes some attempt to subtract out 'natural' change over time.

The difference-in-differences approach can quite easily be combined with a **randomised trial**, although it is usually unnecessary to do so. This is because the difference-in-differences estimate is, algebraically, the same as the difference between the intervention and control group *after* the implementation of the policy minus the difference between the groups *before* the intervention of the policy. In other words, differences between the two groups before the start of the policy are being controlled for. In a randomised trial there should be no such difference so the extra control is unnecessary.

The difference-in-differences design can be combined quite naturally with a **matched area comparison design** and should somewhat improve that design. In this instance the intervention group would be the eligible population within the pilot areas and the control group would be the eligible population within the matched control areas. In both areas, an estimate of before-after change would be made (using either administrative or survey data) and the difference-in-differences estimated.

In instances where the control group is to be selected from eligible non-participants, the choice will often be between a difference-in-differences approach (along the lines of the illustration above) and a **matched comparison design**. The choice between the two depends on the extent to which it is anticipated that the design will remove self-selection bias, with the most likely bias for voluntary active labour market policies being that participants are at a point where they actively wish to find work.

Of the two designs the straightforward difference-in-differences approach is probably the weaker, since the implicit assumption that 'natural change over time' in the numbers with positive outcomes is the same for participants and non-participants is very unlikely to hold if there is self-selection bias (because the participants are a group who are likely to experience greater change than non-participants, with or without the programme).

However, if the matching for a matched comparison design can only be done using administrative variables, then using a *matched* difference-in-differences approach should be an improvement on the straightforward matched comparison group approach. In other words, the two approaches should be combined rather than one design being chosen over the other.

In the instance where more detailed matching is possible, the difference-in-differences approach is likely to be unnecessary, since the baseline differences between the control and intervention groups should be small.

4.5.4 Another use of difference-in-differences

One variation of the approach – which is useful in some instances - would be to use the whole of the eligible population as the intervention group and another, similar, but non-eligible population as the control group. For example, the intervention group for NDLP is lone parents on Income Support, and the control group might be young unemployed women without children within the same area. Under the (strong) assumption that these two groups are seeking the same jobs and experience the same local economic changes, a difference-in-differences approach might be used to identify an NDLP effect.

In practice, this approach is usually considered unworkable because a control group who experience the labour market in the same way as the intervention group is impossible to find.

4.5.5 Examples of the difference –in differences approach

Earnings Top-Up (ETU)

The impact evaluation of ETU, the main design of which is described in Section 4.2, incorporates a simple difference-in-differences estimator. In both the pilot and the comparison areas, administrative data on outcomes was collected both before and after the introduction of the pilots. Change over time will be estimated for both types of area, and the impact of the programme will then be measured by comparing the change in the pilot areas with the change in the control areas.

New Deal for Young People (NDYP)

The impact of the NDYP on youth unemployment was estimated by comparing change in employment rates for young people before and after NDYP was introduced with change over the same period for a slightly older age-group (who were assumed not to be affected by NDYP).

4.6 Cost-benefit analysis

Although not a design approach in itself, the ability to carry out cost-benefit analysis is important within many impact evaluation designs. Cost-benefit analysis attempts to identify all the costs and benefits arising from a programme to provide an overall assessment of its impact. Comparisons can be made to determine if the benefits outweigh the costs and if the benefits net of costs exceed those of alternative schemes. Cost-benefit analysis can, therefore, inform decisions on whether to embark upon or continue a policy or programme, and choices between alternative programmes.

Assessing costs and benefits requires firstly identifying the effects or impact of a policy or programme (which is essentially where the impact evaluation comes in), and then valuing these effects. The results are typically written in monetary values which means that additional positive outcomes (such as additional entries to work) attributable to a programme have each to be assigned a cost.

Placing values on outcomes is clearly problematic: in particular the value attributed is likely to differ depending upon the perspective of the agents doing the cost-benefit analysis. Furthermore, a thorough cost-benefit analysis will need to recognise that benefits have different values depending upon who receives the benefit. However the data requirements for this level of sensitivity will often prove prohibitive.

By way of illustration, the ONE cost-benefit evaluation will attempt to assess the social security savings, taxation gains and wider economic benefits of getting clients into work, set against the costs of the programme. This requires an estimate of the additional numbers of clients leaving benefit for work as a result of ONE (i.e. the impact of ONE). Wider social effects such as those upon crime and health will not be included in the cost-benefit analysis because of the difficulty of measuring any ONE impact in these areas.

As a second example, which illustrates the sensitivity of the findings, cost-benefit analysis of the New Deal for Lone Parents (NDLP) suggested that the prototype programme resulted in economic returns which were slightly less than the cost of the prototype. It was estimated that 20 per cent of the jobs gained by lone parents was as a result of the programme intervention. If that figure had been three percentage points higher – namely 23 per cent – the programme would have had economic benefits equal to its costs.

5 CONCLUSION

This paper is an attempt to give an *overview* of the main evaluation methods used, in particular, within the DWP (formally DSS). However, much of the necessary detail that would be needed to design evaluations of new policies is missing. In subsequent papers of the series the various aspects of evaluation design will be considered in much more detail. In particular, longer papers on process evaluation and impact evaluation will be published; and specific papers on propensity score matching, longitudinal qualitative research, the use and interpretation of qualitative research, and the uses of event history data in evaluation will follow. We recommend that these papers be referred to by any researchers faced with the daunting task of designing the evaluation of a new programme.