CABINET OFFICE

Large Scale Social Experimentation in Britain:

What Can and Cannot be Learnt from the *Employment Retention and Advancement* Demonstration?

David H. Greenberg,
University of Maryland, Baltimore County, United States

Stephen Morris
Department for Work and Pensions

Government Chief Social Researcher's Office
Occasional Papers Series No. 3

The authors are thankful for helpful comments from Jane Hall, Sue Duncan, Annette King and Len Davies on earlier drafts of the paper. Work on the paper was conducted while both authors were working in GCSRO. We are grateful to our colleague Phil Davis for his support and advice over this period.

The report is No. 3, GCSRO Occasional Papers series.

November 2003
Government Chief Social Researcher's Office

ISBN 0 7115 0450 4

CONTENTS

EXECUTIVE SUMMARY

The Employment Retention and Advancement (ERA) Demonstration project is a major new welfare-to-work social experiment, the largest random allocation evaluation ever mounted in Great Britain. This paper draws on experience gained in designing the ERA Demonstration to explore the strengths and limitations of social experimentation for policy evaluation and analysis, and to highlight some of the key issues that need to be considered in designing random allocation experiments.

Testing new interventions through a social experiment

The ERA Demonstration project will begin towards the end of 2003. The Demonstration will test a package of new services and financial incentives that aim to encourage groups on the margins of the labour market to obtain a job, retain work and advance in employment. Specifically, a new type of personal adviser service – the Advancement Support Adviser – will be tested alongside two new financial incentives: a retention and advancement bonus and a training bonus. The effectiveness of these new services and incentives will be compared to the effectiveness of existing services, notably the New Deal initiatives and financial incentives such as tax credits.

The new services and incentives developed through the ERA programme will be thoroughly tested in six areas of the country.

There will be three target groups for the Demonstration: those eligible for the New Deal for Lone Parents (NDLP) and the New Deal for Long-term Unemployed (ND25+), and lone parents working part-time and claiming the Working Tax Credit (WTC). The centrepiece of the evaluation will be an impact study based on an experimental design. Individuals eligible for the ND25+ and NDLP will be randomly allocated to either continue with the New Deals (thereby serving as a control group) or to the ERA programme (thereby serving as a programme group). Similarly, lone parents working part-time and claiming the WTC will be allocated at random to either continue claiming the credit (thereby serving as a control group), or to receive ERA services and incentives in addition to the WTC (thereby serving as a programme group). Impacts will be measured as the difference in mean outcomes (e.g. earnings) between the treatment and control groups.

Random allocation

Random allocation is adopted to estimate the impact of the ERA programme because it provides unbiased or 'internally valid' estimates of the programme's impact. It does so because random allocation ensures that the only differences between programme and control groups at the point of randomisation are random differences – in other words, it ensures that there are no systematic differences between the two

groups, and consequently they are statistically equivalent. Counterfactual estimates of programme outcomes (that is, the mean value of outcomes that would have prevailed for the programme group had they not received new services and incentives) can be estimated from the control group. In the absence of the programme, the only difference between the mean values of outcomes for individuals in the control and programme groups are differences that occur at random. As a result, counterfactual estimates of programme outcomes, derived through a control group constructed at random, are considered 'unbiased'.

A range of alternative quasi-experimental approaches to measuring the effectiveness of the ERA services and incentives can potentially be used instead of random allocation. For example, estimates of programme impacts can be derived from simple 'before and after' type estimators. Alternatively, counterfactual samples can be selected from carefully matched comparison or control areas. Those eligible for a programme who fail to join it can also be sampled and used to construct counterfactual estimates. Some form of matching, such as that based on propensity scores, can be used to improve quasi-experimental estimates of programme impacts. Despite these refinements, all of the quasi-experimental alternatives to random allocation possess substantial drawbacks. The crux of the problem centres on the inability of quasi-experimental methods to deal convincingly with the problem of unobserved selection bias.

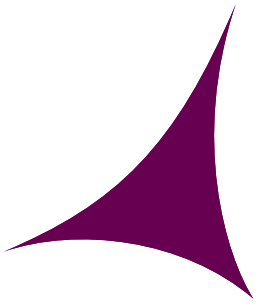
Some design issues in social experimentation

Notwithstanding the benefits of an experimental design, significant barriers exist to the proper implementation of random allocation. Moreover, there are clearly instances where random allocation is unsuitable on ethical grounds.

The twin problems of 'crossovers' and 'contamination' provide appreciable challenges to the designers of social experiments. Crossovers occur when individuals are no longer allocated to programme and control groups by chance alone, and some systematic component enters into the process of allocation. Contamination occurs when individuals assigned to the control group inadvertently receive services or treatments intended for programme group members.

The ERA experimental design seeks to limit the potential for crossovers and contamination to occur, through ensuring that, where possible, programme services and incentives are delivered by a separate group of staff. Furthermore, technical advisers will be on hand to ensure that frontline staff observes random allocation protocols and that administrative records are kept so that it is clear whether a given individual is a member of the programme or control group. A centrally-administered random allocation algorithm ensures that both administrators and customers are *unable* to 'game' the allocation process.

One of the key issues in ensuring that social experiments produce useful findings is to consider carefully the selection of localities (experimental sites) where the experiment will be implemented. Ideally, experimental



sites would be selected at random, but this is seldom feasible and was not an option for the ERA Demonstration. Instead, six experimental sites were selected from across Great Britain on the basis of a number of criteria. These included the need to avoid Jobcentre Plus[1] districts engaged in major administrative reorganisation, the need to obtain a reasonable geographical spread of sites and the need to be able to select samples of sufficient size in each site, such that programme impacts might be detected at site level.

In common with many quasi- and experimental evaluations in Great Britain, the ERA Demonstration is largely reliant on survey data to measure outcomes. The larger the size of survey samples, the smaller the impacts that can be detected. The problem is that it is expensive to collect data from survey samples. The ERA Demonstration project used the concept of *Minimum Detectable Impact* to identify the most appropriate trade-off between cost and sample size.

Random allocation designs rely on computing the difference between average values for programme group outcomes (e.g. earnings) and averages control group outcomes, in order to estimate the impact of the programme or intervention under investigation. In many social experiments, however, simple experimental comparisons are not sufficient to address the full range of questions that evaluators wish to consider. In the case of the ERA Demonstration, one of the key issues is whether the programme leads to improvements in hourly wages and increased wage progression among those

in the programme group. Because only a fraction of the programme and control groups will enter employment and thereby record hourly wages, and it is anticipated that the process of obtaining work by members of the programme group will be influenced by ERA services and incentives, it is highly likely that a simple comparison of wage rates between employed programme and control group members will not yield unbiased estimates of programme effectiveness. For this reason, quasi-experimental methods will be required in addition to simple comparisons of programme and control group outcomes.

What will not be learnt from the ERA Demonstration

Social experiments seek to answer questions about causality and the impact of programmes or interventions. There is a range of questions of interest to policymakers and evaluators, however, which social experiments can either not address at all (because they are not designed to do so) or that can only be addressed with specific modifications to the experimental design. But such modifications often render the practical implementation of experiments problematic.

One of the main charges levelled against experiments is that they fail to provide an explanatory account of the processes that give rise to observed programme impacts. This limitation is frequently termed the 'black box problem'. For example, the ERA Demonstration involves the delivery of both caseworker services and financial incentives as a single package. The experimental design – the allocation of participants to a single programme group or to a control group – does not allow *separate* experimental estimates of the impact of Advancement Support Adviser (ASA) services and the

[1] The ERA Demonstration project is to be delivered through Jobcentre Plus.

separate impact of financial incentives. In order to address the issue of the relative effectiveness of different elements of the ERA programme, more complex, differential, randomised designs are required. These designs require both larger sample sizes to make multiple comparisons and place a greater administrative burden on frontline staff, consequently increasing the likelihood of administrative error. For these reasons, a differential design for the ERA programme was rejected, despite the analytical gains that can result from such designs. As a result, the evaluation of ERA relies heavily on a non-experimental, observational process study to uncover evidence of the separate contributions that different components of the programme make to programme impacts, should these impacts actually materialise.

A critical issue in evaluation is that of 'external validity' – the extent to which estimated programme impacts can be generalised to different locations and populations, to different time periods and to different variants of the programme being studied. Generalisability is an issue for all forms of evaluation, including social experimentation. Results from an experiment might not hold at different time points and in different geographical localities. Experimental impact estimates are usually derived from the context of a pilot or demonstrations limited to a particular set of areas and are thus smaller in scale than in a national programme. As a result, it may be problematical to infer the impact of a national full-scale programme from a smaller-scale experiment. Furthermore, substitution effects, Hawthorne effects, entry effects and general equilibrium effects may all limit the capacity to draw generalisable estimates of programme impacts from a single experiment.

Conclusions

The ERA Demonstration illustrates both the strengths and weaknesses of social experiments in evaluating social programmes. For evaluating ERA, and a wide variety of other social policy interventions, an experimental design is superior to alternative designs that might be used instead – for example, a 'before and after' comparison, matched sites, or a participant/non-participant comparison. It will provide greater assurance of internal validity, while being no more costly or time-consuming. However, this does not mean that experimental designs are always superior for evaluating all social policies; just that experiments are often advantageous, and that random allocation clearly is the best approach for evaluating ERA. Quasi-experimental methods may be less expensive and less time-consuming than random assignment for evaluating existing programmes. Moreover, occasionally there are ethical reasons for not using random allocation. Nonetheless, if implemented and run properly, an experimental design will almost always provide greater internal validity than alternative approaches.

No single evaluation design can answer all the questions about a specific social policy that are of interest, and random allocation is no exception. Sometimes, however, certain design modifications can be made that can help address certain issues. For example, although ultimately not adopted, consideration was given to using a differential experimental design for the ERA Demonstration in order to determine whether the impact of combining financial incentives with services is greater than the impact of financial incentives alone. Other limitations of a single evaluation design can be at least partially overcome by combining several different approaches. For example,

quasi-experimental econometric methods will be required to examine certain issues concerning ERA's impact on advancement, while a process study will be used to help determine the context and the manner in which ERA services were delivered.

There are certain important questions that no combination of evaluation methods can definitively address, however. For example, neither experimental nor non-experimental methods will be able to provide more than limited information about which specific components of ERA are most or least effective – the so-called 'black box problem'. In addition, once findings from the ERA Demonstration become available, uncertainty will inevitably remain about their 'external validity' – that is, the extent to which they can be generalised to different locations and populations and to different time periods; whether they are subject to scale bias, general equilibrium wage effects, substitution effects and/or Hawthorne effects; and whether entry effects might occur if ERA is rolled out nationally that did not arise during the Demonstration – regardless of the combination of experimental and non-experimental methods that were used to obtain them.

1. INTRODUCTION

Social experiments are field trials that randomly allocate individuals to programme and control groups for the purposes of evaluating new social programmes or changes in existing programmes. Since the 1960s, over 200 such experiments have been conducted in the United States (Greenberg and Shroder 1997), by far the largest number in any country. The number of social experiments conducted in the United Kingdom, where at least a dozen such evaluations have been undertaken, is probably second only to the number conducted in the United States. Descriptions of a selection of social experiments undertaken in the United Kingdom in the area of welfare and employment are presented in the Annex to this paper. These experiments are not well known. Many are small and some, for a variety of reasons, did not produce usable findings; others, however, generated useful results.

The advantages and disadvantages, and strengths and weaknesses, of random allocation experiments for evaluating social programmes have been debated for many years (Burtless 1995; Burtless and Orr 1986; Cook and Campbell 1979; Heckman and Smith 1995; Pawson and Tilley 1997). These discussions, however, often tend to be rather abstract in nature and various authors tend to take one side or the other. In contrast, this paper considers both the advantages and disadvantages of social experimentation in the context of a specific random allocation demonstration that will be the largest yet undertaken in the United Kingdom – the Employment Retention and Advancement (ERA) Demonstration. The ERA Demonstration will test services and financial incentives that are intended to help disadvantaged individuals obtain and retain work, as well as helping them advance in employment.

The plan of this paper is as follows. The next section describes the programme that will be tested by the ERA Demonstration and explains why it is important to thoroughly test it before it is introduced nationally. Section 3 describes the random allocation design that will be used to evaluate the ERA pilot programme and some alternative approaches to impact evaluation that might have been used in its place. It then considers why these alternative approaches were not adopted. Section 4 discusses some of the difficult issues that had to be confronted in planning the ERA Demonstration, issues that are typically faced by those attempting to implement social experiments. Section 5 examines the types of information that, while relevant to policymakers, social experiments such as the ERA Demonstration cannot address. Most of these limitations similarly apply to other forms of non-experimental impact estimation. Some conclusions are presented in Section 6.

2. THE EMPLOYMENT RETENTION AND ADVANCEMENT PROGRAMME

Programme objectives

The ERA Demonstration, which will begin in late 2003, aims to test a new policy intended to help those on the margins of the labour market obtain and a job, retain work and advance (Morris *et al*. 2003). The policy combines new and existing services with financial incentives to achieve these goals. The underlying rationale involves the provision of combined pre-employment and in-work support services for those who are initially out of work, and in-work support for those already in low-wage employment. For those who are initially out of work, services are available prior to job-entry for nine months, when the focus is on re-attachment to the labour market combined with a view to longer-term sustainability. For this group, job retention and advancement services will continue to be available for up to two years after entry into work. For those already in low-paid jobs, services intended to encourage job retention and advancement will be available immediately on entry into the programme and will continue for up to 33 months.

The ERA programme's objectives include encouraging both job retention and employment advancement. Job retention is defined as sustained employment in any job of 16 hours a week or more. The objective is to prevent breaks from occurring in an individual's work record, thereby helping them avoid time spent claiming benefits. To the extent possible, the programme will attempt to locate jobs for individuals that provide opportunities for advancement.

For this reason, one of the key features of the ERA programme is assistance with job-to-job moves through the provision of help with job search and help in identifying jobs with opportunities for advancement.

The concept of advancement is multifaceted and arriving at a neat definition is more challenging than with job retention. Individuals who increase their annual earnings might be considered to have advanced. However, increases in earnings can occur through either more hours worked or improvements in hourly wages, or both. For this reason, an improvement in the hourly wage is clearly an important indicator of advancement. But hourly wages can rise, while overall earnings fall as a result of a simultaneous reduction in hours. Thus, both hourly wages and earnings need to be considered together in determining whether an individual has advanced.

Apart from earnings and wages, there are other conditions of employment that should be taken into account in assessing the extent of advancement – for example, whether an individual has a supervisory role as part of their job, whether they enjoy pension benefits provided by their employer, or whether they have access to paid holidays and other such fringe benefits. Moreover, an individual's subjective assessment of their job also needs to be taken into account. The evaluation of ERA services requires research instrumentation of sufficient scope to assess the impact of the ERA programme on these diverse aspects of advancement.

Target groups

The ERA Demonstration is designed to test new services on three target groups that were chosen for two reasons: they were viewed as the groups most likely to benefit from the programme, and individual members of the groups can be readily identified and located from administrative records.

The three target groups are: those eligible, and in most cases required, to join the New Deal for Long-term Unemployed (ND25+); those who choose to enter the New Deal for Lone Parents (NDLP); and lone parents working part-time and claiming Working Tax Credit (WTC). Appreciable numbers in these groups are known to encounter problems in obtaining, retaining or advancing in work, or in all three. The two New Deal groups will become eligible to enter the ERA Demonstration at the point they would normally enter the New Deal. They will be identified through records held by Jobcentre Plus. Working lone parents will be eligible for entry into the Demonstration as long as they are working part-time (between 16 and 29 hours per week) and are claiming WTC (which replaced the Working Families' Tax Credit in April 2003). They will be identified through records held on the WTC administrative database.

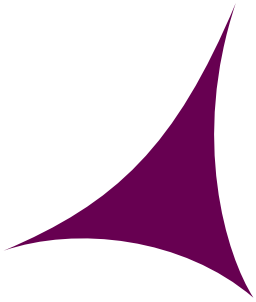
Programme components

The services to be tested through the ERA Demonstration comprise two components: caseworker services and financial incentives. Here, we provide a brief overview of these services. Readers interested in a fuller discussion should refer to Morris *et al.* (2003).

Caseworker services are to be delivered through an Advancement Support Adviser (ASA) who will be located within a Jobcentre Plus office. The ASA and each individual enrolled in the programme will jointly develop an Advancement Action Plan (AAP). For those not in work, the AAP will set out agreed steps that need to be taken in order for the individual to find and retain work, as well as to advance after having obtained work. The initial focus of the plan will be on simple steps that can bring a relatively speedy sense of achievement for individuals. After an initial job is obtained, the plan will focus on the steps needed to stabilise work patterns and then on more ambitious strategies for advancement. For those already in work, the AAP will outline actions intended to encourage advancement from the outset, while not losing sight of the need to maintain job retention.

ASAs will have access to a range of resources in order to help individuals achieve the goals set out in their AAPs. They will be able to broker services from a variety of sources to address specific barriers to employment retention and advancement that individuals might face. ASAs will also have access to an emergency fund. The main resource at their disposal, however, will be two financial incentives: a retention and advancement bonus, and a training bonus. Both bonuses will be used by ASAs to support the objectives outlined in each client's AAP.

A retention and advancement bonus of £400 will be payable to individuals on the ERA programme who work at least 13 weeks during a given 17-week period. Each individual can receive a maximum of six bonus payments totalling £2,400 during the lifetime of the Demonstration. They must be working full-time (that is, working, on average, at least 30 hours a week) in order to qualify for the bonus. The bonus has been structured to encourage steady full-time work



on the basis of the theory and empirical evidence, such as exists, that this is most likely to lead to advancement (Arulampalam and Booth 1998; Campbell and Green 2002). Payments of the bonus will be made at regular meetings between individuals and their ASAs, held every 17 weeks.

The training bonus aims to support training agreed to and set out in an individual's AAP. A bonus of £8 per hour, multiplied by the course length in hours will be payable for successful completion of an agreed training course, up to a maximum cumulative total amount of £1,000 for each individual, over the lifetime of the ERA Demonstration. In addition, a fund of £1,000 to pay for course fees will be available for each individual.

The need for a rigorous evaluation of the ERA Demonstration

New policies that aim to improve levels of job retention and employment advancement, such as the ones described above, are certain to be expensive. The opportunity costs of diverting resources toward such services are substantial. Moreover, although ERA is designed to be as consistent as possible with theory and evidence concerning what is most *likely* to be effective, given the lack of existing knowledge in the UK about how retention and employment can be improved, there is no way of knowing in advance that ERA services will actually prove effective. It is therefore extremely important that policymakers have information as to whether ERA services can achieve their objectives at a reasonable cost, before a decision is taken to introduce the programme nationally. Only in this way can policymakers be sure that the substantial resources required to fund the ERA programme are being used in a way that is productive.

To obtain the needed information about the effectiveness and costs of ERA, a policy demonstration pilot is being carried out. The ERA Demonstration will run in six geographical areas (known as programme sites) over a three-year period. Because the evaluation of the effectiveness of the pilot programmes will not be completed for two more years, it will be five years before policymakers will have all the information the pilot will provide at their disposal, although preliminary findings will be available a couple of years after the pilots begin in 2003. Such delays are, of course, frustrating. However, many of the phenomena policymakers are interested in, particularly job retention and employment advancement, can only be measured successfully over an extended period of time.

As indicated in the Introduction, the centrepiece of the Demonstration will be an impact assessment in the form of a randomised social experiment. Estimates of programme effects that are based on a random allocation design are to be combined with other methods, including a process study that explores the causal mechanisms at work through the programme, a thorough cost study and a cost–benefit analysis. A multi-method approach provides the best chance of successfully addressing a wide range of questions about the effectiveness of the ERA programme. For example, did the policy generate the intended impacts? What was the nature of the causal mechanisms that generated these impacts? And, did the benefits of the programme outweigh the programme's net costs?

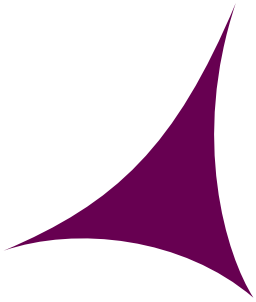
3. WHAT IS RANDOM ALLOCATION AND WHY IS IT BEING USED TO EVALUATE ERA?

Most methods for establishing a causal relationship between a new policy (or a change in an existing policy) and observed changes in an outcome of interest to policymakers, involve attempts to determine counterfactual outcomes. In other words, they attempt to establish what would have occurred had the new policy not been introduced, or a change to existing policy not been brought about. These approaches examine whether changes in outcomes of interest to policymakers can be attributed to the policy by comparing average outcomes for the individuals affected by the programme with average counterfactual outcomes.

One way of establishing a counterfactual is through random allocation. Individuals who are eligible for a programme are assigned to either a programme group or a control group by chance alone. In the case of ERA, individuals in the target groups will have an equal chance of being assigned to the programme group or the control group. Those assigned to the programme group will have access to ERA services and financial incentives, while those in the control group will not, but will instead be eligible for all existing non-ERA assistance (e.g. the New Deal and Jobcentre Plus services as well as the WTC). Thus, in the ERA evaluation, the control group will represent the counterfactual.

Programme impacts

As already suggested, much of the evaluation of a random allocation experiment is based on comparisons of average outcomes between programme and control groups. For example, one of the key anticipated benefits from ERA is that, as a result of greater job retention, members of the programme group will work more weeks than their counterparts in the control group. The extent to which this occurs can be readily measured by simply subtracting weeks worked by members of the control group after random allocation, from weeks worked over the same time period by members of the programme group. This difference, which, typically in practice, is statistically adjusted through the use of regression analysis, provides an estimate of the programme's 'impact'. Numerous different types of ERA impacts are potentially of interest and can be estimated in a similar fashion. Examples include impacts on hours worked per week, earnings, benefit receipt and health status. In addition, because members of the programme and control groups will receive some similar services (for example, job search assistance and training), it is also of interest to determine whether ERA has an impact on the amount of such services that are received. Finally, the Government will incur costs in serving both groups, but because ERA will provide new services and financial incentives, it is anticipated that the costs of serving the programme group will be larger. This can be measured by estimating ERA's impact on costs.



Alternatives to random allocation

The purpose of estimating programme impacts is to determine the difference made by the programme being evaluated. Doing this is only possible by comparing various outcomes with the programme (e.g. weeks worked, earnings, costs, receipt of services, etc.) and without the programme. Random allocation provides one means for making such a comparison, but other types of comparisons, which are known as 'quasi-experimental comparisons', are also possible. In general, quasi-experimental comparisons are considered inferior to experimental comparisons (Boruch 1997; Burtless and Orr 1986; Cook and Campbell 1979; Orr 1999; Purdon 2002; Shadish, Cook and Campbell 2002; among others), although, as discussed later, social experiments are subject to certain shortcomings of their own. As will be seen, the key problem with non-experimental comparisons is that there is no way to guarantee that the groups being compared do not differ systematically from one another for reasons that have nothing to do with the programme being evaluated. If they do differ, the comparison cannot be said to be 'internally valid' (Campbell and Stanley 1963; Cook and Campbell 1979) or, in other words, the comparison may be biased.

One alternative to random allocation is a 'before and after' comparison. For example, ERA could be evaluated by collecting outcome data on members of target groups within the six pilot-site areas for several years prior to the project being implemented. Once ERA had begun, similar data would again be collected on members of the target groups in the same areas. The second set of individuals would serve as the programme group and the first as a comparison group. (In the literature on evaluation design, the term 'control group' is often reserved only for those non-programme groups that are created through random allocation, a convention that we also adopt.) Because the membership of the target groups inevitably changes with the passing of time – for example, members of the New Deal groups obtain jobs and lone parents on WTC get married or leave employment – the two sets of individuals on whom data were collected would not be identical, although there would inevitably be some overlap.

One problem with this approach is that, after ERA had begun, a large fraction of WTC lone parents and a small fraction of NDLP lone parents and those in the ND25+ target group in the programme sites would probably elect not to participate in ERA. However, there is no way to positively identify which individuals in the comparison group would have made the same decision not to participate had they been given the opportunity to make it. Thus, to maintain comparability between the two groups, data would need to be collected on everyone in both groups, not just those individuals desiring to participate in ERA. A more serious problem is that changes may occur over time that do not result from ERA, but nonetheless affect the programme group. For example, the economy could change and, as a result, some or even all of any measured differences between the programme and comparison groups in weeks worked or earnings may not be attributable to the ERA. It can be extremely difficult to determine the portion of the difference that would be attributed only to the ERA programme.

Another alternative to random allocation is a comparison or 'matched sites' approach. Under this method, outcome data would be collected for members of the target groups in both a set of programme pilot sites and a set of non-pilot sites. Individuals in the first set of sites would serve as the programme group and those in the second set of sites as the comparison group.

As with the 'before and after' comparison, in the case of a programme such as ERA, some members of the target groups in the programme sites would elect not to participate, but there is no way to identify with certainty those individuals in the comparison sites who would have made the same decision not to participate, had they been given the choice. Thus, to maintain comparability between the programme and comparisons groups, data would need to be collected on both participants and non-participants in the programme sites. More seriously, there would almost certainly be differences in outcomes between the programme and comparison groups for reasons having little to do with ERA. For example, the pilot and non-pilot sites will differ in terms of the characteristics of their client populations, the quality of staff at the Jobcentre Plus offices, and the characteristics of the local economies. To some extent, these differences can be controlled for statistically by carefully matching the pilot and non-pilot sites, but as discussed below, this is rarely sufficient (Heckman, Ichimura and Todd 1997). As a practical matter, the matching will be imperfect because, while the sites will differ from one another along a large number of different dimensions, only a limited number of criteria can be used for matching (Friedlander and Robins 1995; Hollister and Hill 1995). The problem also becomes less severe as more programme and comparison sites are added, both because more criteria can then be used for matching sites and because remaining differences between the two sets of sites tend to wash out. Indeed, with a sufficient number of sites, the sites themselves can be randomly assigned. However, as the number of sites grows, so will evaluation and programme costs, because the programme must be administered and data must be collected in more locations. Moreover, even if sites are randomly assigned to programme and control status, it is doubtful that there will be a sufficient number of sites to assure that the two groups of sites do not differ in some unobserved way.

One possibility for considerable cost saving with the 'matched area' design is to use data that are already being collected on members of the target groups in the comparison sites, rather than introducing new surveys in these areas. Such data, for example, are currently collected for administering and assessing New Deal programmes and the WTC. For the purposes of the ERA evaluation, however, it will be necessary to survey members of the programme group for several years after they enter the programme, regardless of whether they remain on benefit or not. Similar information would also be needed on members of the comparison group, but existing data sources do not provide comprehensive long-term information of this sort.

As a final alternative to random allocation, those who *do not* participate in the pilot programme could be used as a comparison group, where such a group is located in the same sites, at the same time, as those who *do* participate. This 'participant/non-participant comparison' has the major advantage of not subjecting comparisons between the two groups to changes that occur over time or to differences between sites. The problem with this approach is in locating non-participants who are comparable to those in the programme group. In important respects, those outside the ERA target groups (for example, low-wage married women who are working part-time or men on Jobseeker's Allowance who have been unemployed for fewer than 18 months) are unlikely to be very similar to those within the target groups (for example, WTC lone parents or men on

New Deal 25 plus). Another possibility is to use people in the target groups who *choose* not to participate in ERA. This is especially feasible in the case of the WTC lone-parent target group, as a substantial fraction of those in this group probably will opt not to participate. However, one would suspect that those who do decide to participate would differ systematically from those who choose not to participate – for example, in terms of drive and motivation. If they do, a 'selection problem' is said to exist. In other words, those selecting themselves as participants differ from those who select not to participate.

A number of statistical techniques have been developed to attempt to control for these differences. Perhaps, the most popular are various forms of statistical matching. These procedures essentially involve matching individuals in the programme group statistically with individuals in the comparison group on the basis of various observed characteristics such as age, sex, race, education, and work experience. In some cases a composite index is estimated, based on the characteristics of members of both the programme and comparison groups, which estimates the probability of participation in the programme for both groups, known as the propensity score (Rosenbaum and Rubin 1984). Matched samples can then be constructed by matching on the propensity score. Matching techniques could potentially not only be used to evaluate ERA through 'participant/non-participant' comparisons, but also in the case of 'before and after' comparisons or 'matched site' comparisons. They could also be used to limit the analysis to only those individuals who choose to participate in the ERA programme. However, there is usually no way to know if the matching procedure has succeeded or not. Moreover, when it has been possible to test the success of matching procedures, the results have not generally been very encouraging (Bloom *et al*. 2002; Friedlander and Robins 1995; Glazerman, Levy and Myers 2002; Heckman, Ichimura and Todd 1997). This is probably because it is not usually feasible to match on characteristics, such as drive and motivation, which are not readily measured but nonetheless may influence programme participation decisions.

The major reason for using random allocation to evaluate ERA is that numerous studies have now accumulated (Bloom *et al*. 2002; Fraker and Maynard 1987; Friedlander and Robins 1995; LaLonde 1986; LaLonde and Maynard 1987; among others) that demonstrate that random allocation produces considerably more reliable estimates of programme impacts than any other method of estimating impacts, including those outlined above. Random allocation is simply not subject to the sorts of problems facing the alternative methods. The reasons why are discussed next.

What are the advantages of random allocation?

When both the measured and unmeasured characteristics of individuals in the programme group are statistically equivalent to the characteristics of groups or individuals acting as a counterfactual, but for the fact that the former are exposed to the programme or policy being evaluated, such a comparison can be said to be 'unbiased', 'internally valid' and free from 'selection bias'. The major advantage of random allocation over the alternative approaches discussed in the previous subsection is that, as a method for determining the impact of a policy or programme, programme estimates *do not* suffer from an 'unknown degree of bias'

(Burtless and Orr 1986: 609). Central to the assumptions frequently underlying most non-experimental methods are those that must hold for internal validity to be achieved or, viewed slightly differently, to avoid selection bias. As has been discussed, for the impact of ERA services, which are determined by comparing outcomes for a comparison group with those for a programme group, to be 'internally' valid and free from selection bias, the two groups have to be statistically equivalent.

Random allocation is the method most able to ensure 'statistical equivalence' between a programme group and a counterfactual, or control group, because both groups are created by chance alone. Comparisons of average outcomes between the programme and control groups are internally valid, because the only systematic difference between the two groups is that the programme group is exposed to the new policy or programme, but the control group is not. Certain assumptions have to hold in order for impacts estimated through random allocation to be considered 'internally valid'. (For instance, the process of random allocation *must not* affect the behaviour of those assigned to the control group so that they act as a genuine counterfactual. If the process of random allocation does affect the behaviour of individuals in the control group, for example, if they are stimulated into accessing services they otherwise would not receive, programme impact estimates will be biased.) These assumptions, however, are fewer in number and, if invalid, are potentially less substantial in their implications than the assumptions that need to hold for non-experimental methods to be considered free from 'selection bias'. It should be noted that none of this discussion obviates the need for social experiments to be properly designed and run, if the benefits of random allocation are to be achieved.

Random allocation possesses an additional advantage over non-experimental methods in that the results from social experiments are relatively simple to explain to non-technical audiences (Burtless 1995; Orr 1999). The assumptions underpinning most non-experimental methods, which are necessary for the estimation of unbiased findings, are complex and difficult to understand, requiring familiarity with statistical and econometric concepts not usually found among policymakers and other users of research. Social experiments, however, provide policymakers with accessible, easily interpreted estimates of average policy or programme impacts. They do so without the need for making a range of complex assumptions, which could cause large biases in estimates of programme impacts were they not to hold.

Should random allocation always be used?

There are circumstances, nevertheless, when it is inappropriate to use social experiments for impact analysis. Discussions about these circumstances can be found in Cook and Campbell (1979), Orr (1999), Rossi, Freeman and Lipsey (1999) and elsewhere. The main issues are briefly discussed in this subsection, as well as later in this paper.

As Weiss (1998) has pointed out, there is a right time and there is a wrong time to evaluate. For social experiments, or any other evaluation design to yield useful results, the programme or policy under investigation needs to be stable and, therefore, not prone to substantial alteration or major

reorganisation (Freeman, Rossi and Lipsey 1999). In essence, the causal mechanisms, the effects of which are under investigation, need to be established. Consequently, it is often advisable for evaluators to conduct an evaluability assessment (Weiss 1998) prior to subjecting a policy or programme to a full experimental impact study.

Evaluations of small-scale pilot demonstrations, including random allocation evaluations, may not provide very useful information about the potential effects of policy changes that are expected to have large entry effects, or large effects on community attitudes or the macro-economy (Garfinkel, Manski and Michalopoulos 1992). These effects are discussed in some detail in Section 5. However, the basic idea can be illustrated by recent welfare reform in the United States, which seems to have resulted in sea changes in attitudes among both the public assistance client population and the caseworkers who serve this population. These changes in attitude appear to have discouraged entry into public assistance programmes and encouraged exit from these programmes. This, in turn, may have resulted in substantial increases in the supply of workers seeking low-paying jobs, possibly keeping wages in such jobs lower than they otherwise would have been. Such effects are unlikely to result from small-scale pilot programmes, which are limited to half the target population in a limited number of sites, and will therefore be missed by evaluations of these pilot programmes (Moffitt 2002). Although there is no reason to anticipate that the ERA programme would have large effects of this sort, even if rolled out nationally, the potential for some of these effects to occur, albeit on a smaller scale, cannot be dismissed. As discussed in Section 5, a potential limitation of the ERA evaluation is the possibility that such effects *might* be missed.

There is also a range of circumstances where it might be considered unethical to use experimental methods; however, in many circumstances social experiments are entirely ethical. It is commonly argued, for example, that random allocation is unethical because members of the control group are prevented from accessing services available to the programme group. In other words, random allocation ensures that the programme and control groups are statistically equivalent; yet they are treated differently. Such a charge can be easily dismissed. Put simply, the differential treatment is ethical as long as it is impossible to know in advance whether the services to which the programme group has access are beneficial and that these benefits are generated at a reasonable cost. The only way to determine this is to conduct an impact study and, as previously discussed, an impact study is only possible if a counterfactual is established and this, in turn, usually requires the exclusion of some individuals from programme services.

Nonetheless, there are certain circumstances under which the use of random allocation has been clearly established as unethical. For example, the programme group should not be exposed to an intervention known *a priori* to involve some 'positive harm'. Likewise, social experiments should not be used when the design requires the withdrawal of an existing 'good' from the control group. There is no reason to anticipate that ERA will either harm members of the programme group or result in any losses of an existing 'good' to controls. Moreover, the impacts of ERA on participants are *not* known in advance. Thus, there is no ethical reason why the impact of ERA services should not be evaluated through random allocation.

Other objections or limitations of social experiments that are commonly put forward can be addressed by making sure that

random allocation is properly designed and implemented. Two further objections, however, are that social experiments are disproportionately costly and that it takes too long to obtain usable results from them. These criticisms can have considerable validity under certain very specific circumstances. Imagine, for example, that ERA had been operating during the previous five years and that outcome data were collected on programme participants during this period. Further, imagine that similar data were available for a non-randomly assigned comparison group. If a decision was made today to conduct a non-experimental evaluation, a comparison between the two groups could be made rather quickly and inexpensively. However, if the decision was to conduct a random allocation evaluation instead, the previously collected outcome data could not be used. Instead, random allocation would have to take place and then new outcome data would need to be collected over several years. Only then would it be possible to compare the programme and control groups. Thus, there would be a considerable delay before results were available. Note, however, that the experimental evaluation would not necessarily be much more expensive. In both instances, the programme would have to be run, outcome data would have to be collected, and the data would have to be analysed. The only difference is the cost of implementing and monitoring the random allocation process itself, but this is generally relatively inexpensive.

In fact, of course, ERA has not yet been implemented. When it does start, an evaluation of it will require the collection of outcome data and the analysis of these data. Thus, it should not be more time-consuming than the various alternatives to randomisation described above. In fact, if the 'before and after' design, previously described, was used instead of random allocation and appropriate data did not already exist for the comparison group, it could actually take longer to obtain evaluation results. Moreover, except for the costs of implementing and monitoring randomisation, it should not be more expensive. Indeed, it could be less expensive than the 'matched sites' design if new data had to be collected on individuals in comparison sites, because surveys would have to be conducted in more areas. Furthermore, several of the alternatives to random allocation that were described above, require that data be collected on individuals who are not interested in participating in ERA, as well as those who do wish to participate. A random allocation evaluation, in contrast, only requires data on those who express interest in participating in the programme being evaluated.

4. SOME DESIGN ISSUES

In this section, we discuss some design issues that confront most social experiments. Several of these issues must also be faced by non-experimental evaluations. We describe how the ERA evaluation design aims to mitigate some of these problems, bearing in mind that there is no way to eliminate some of the problems entirely.

Contamination and crossovers

A key issue in any social experiment is ensuring that random allocation actually occurs. This means that whether a given individual is in the programme group or the control group actually depends on chance alone, and that members of each group only receive the services for which they are eligible. Violation of the first of these provisions is said to result in 'crossovers' and violation of the second in 'contamination', both of which compromise internal validity. Crossovers decrease the statistical equivalence between the two groups, while contamination biases estimates of programme effects downward because it reduces differences in the extent of new treatment received between the programme and control groups. Careful planning can greatly reduce the possibility of contamination and crossovers.

One way in which ERA has been designed to reduce contamination, is by ensuring that members of the programme group will be served by a completely new group of caseworkers, the Advancement Support Advisers (ASAs), working entirely separately from caseworkers delivering services to control-group members. In addition, a technical adviser will be assigned to each site. Technical advisers will be responsible for monitoring the random allocation process, ensuring its integrity and making sure that members of the programme and control groups receive only those services to which they are allocated.

Crossovers can result from poor administrative record keeping. They also sometimes occur when programme administrators feel that individuals who are randomly assigned to the control group would be better off in the programme group, or vice-versa.

Random allocation is usually achieved through the application of a random allocation algorithm – a statistical process that ensures programme and control groups are created at random or very close to random. Several different designs for a random allocation algorithm can be used to minimise the possibility of crossovers. The alternative that seems to make the most sense for the ERA pilot is to establish an algorithm comprising a sequence of blocks of a random length for the programme sites. To illustrate, the sequence of blocks under this approach would look something like this:

PPCC, CP, PCPCPC, CPCP, CCPP, CPPCPC, CP, . . . ,

where 'P' represents an allocation to the programme group and 'C' allocation to the

control group. For example, using the illustrative sequence of blocks appearing above, the eighth individual to be assigned would be allocated to the control group.

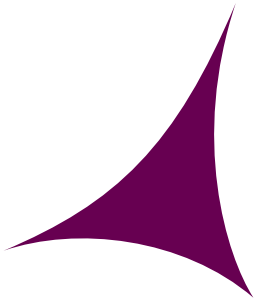
Both the ordering of the Ps and the Cs within each block and the length of each block would be determined randomly, but the number of Ps and Cs within each block would be equal. Because neither Jobcentre Plus staff nor members of the ERA target population will have knowledge of the sequence of block lengths, or the sequence of Ps and Cs within each block, they cannot know in advance the group to which the next individual who enters the study would be assigned. Thus, it will be virtually impossible for either to manipulate the allocation process.

The plan is for the same sequence of blocks to be used to assign individuals randomly from all three target groups. To illustrate, imagine that the first three individuals who are randomly assigned at a particular site are from the ND25+ target group, the next two are from the NDLP target group, and the next two are from the ND25+ target group. Using the illustrative sequence of blocks appearing above, four of the five ND25+ individuals would be assigned to the programme group, while both of the NDLP individuals would be assigned the control group. Given the 'law of large numbers', however, it is likely that by the end of the random allocation process spanning 12 months, after hundreds of individuals from each group have been randomly allocated, the numbers assigned to the programme and control groups within each target group at each site will be approximately in balance. It is unlikely, however, that exact 50:50 ratios will be obtained.

Representativeness of sites

Pilot studies that are used to test a new policy or programme are often conducted at several different sites to determine whether the policy being tested can succeed under a variety of conditions. A key issue is whether the sites selected are sufficiently representative of the population of sites as a whole, so that the findings from the experiment provide information on what would happen if the tested programme were rolled out nationally. This is one of a number of issues concerning the 'external validity', or generalisability, of an evaluation. Several of the other concerns about external validity are discussed in Section 5.

Ideally, in running a pilot study, a large number of programme sites would be selected at random, with the target population at these sites randomly allocated to programme and control groups. In such a situation, estimating the impact of the programme on data pooled from across the sites, and then calculating the confidence interval of that estimate, would enable the evaluator to assess the precision of the impact estimate for the entire target population. In reality, however, random selection of a large number of sites is seldom possible for any type of evaluation, experimental or non-experimental, not least due to the associated cost. Moreover, it is often difficult to find sites that are free from other piloting activity targeting the same or similar groups of people. As a result, most evaluations cannot provide estimates of impact parameters for an entire target population on the basis of their design alone. In our view, however, this does not fatally undermine the usefulness of social experiments. Substantial benefits to policy can still accrue from impact estimates that indicate whether or not a policy has an



impact on members of the target population who reside in a diversity of settings.

Usually then, the evaluator must accept a second-best solution and select experimental sites purposively, rather than randomly. To achieve a selection that is diverse, as well as meaningful, it is important that this purposive sampling of sites is undertaken very carefully. In addition, evaluators very often need to take administrative and political considerations into account in their selection of sites.

The budget available for the ERA Demonstration programme is sufficient to test ERA services in six experimental sites, where each 'site' is equivalent to a Jobcentre Plus district. In selecting the six experimental sites, an attempt was made to avoid areas that at the time the ERA Demonstration began would be in the process of implementing the very substantial administrative reforms required by the introduction of the new Jobcentre Plus service model. However, the aim of the demonstration is to compare ERA services with those provided through the new Jobcentre Plus model. For these reasons, an effort was made to select districts that were due, as far as possible, to have been operating Jobcentre Plus for at least six months prior to the scheduled ERA start date, in order to give the new Jobcentre Plus regime time to bed-down and stabilise. According to the Department for Work and Pensions' (DWP) Jobcentre Plus rollout plan, there were 25 potential sites, at the time localities were to be chosen, that were scheduled to introduce Jobcentre Plus at least six months prior to the launch of the Demonstration and thus satisfied this criterion.

Several other features of the potential pool of Jobcentre Plus districts influenced the selection of sites for the ERA Demonstration. First, it was important that each of the selected sites contained a sufficient number of people in each of the three target groups so that it will be possible to detect programme impacts at the site with reasonable levels of statistical significance and sample power. Second, a sufficient number of members of key subgroups (for example, ethnic minorities) needed to reside in the selected sites so that, after pooling across the six sites, it will be possible to estimate subgroup impacts with an acceptable level of statistical precision. Third, it was important to ensure reasonable regional diversity – for example, to make sure that no two sites were selected from the same geographical region and that there is an even balance between urban, semi-rural and rural sites. Based on these considerations, six sites were selected for the ERA Demonstration. These are East London, Manchester, Gateshead and Tyneside, Derbyshire, south-east Wales and the Scottish counties of Renfrewshire, Inverclyde, Argyll and Bute. The projected target group size estimates for two of these sites are smaller than ideal. It may, therefore, be necessary to extend the period during which individuals can join the programme in these sites to ensure samples of an adequate size.

Determining sample size

Like many UK evaluations, the ERA Demonstration will be largely reliant on surveys to measure programme impacts. This is in contrast to the situation in North America, where welfare-to-work demonstrations are typically able to rely more heavily on administrative data sources in measuring programme impacts. While surveys are able to collect data on a wider

range of outcomes than is available from administrative sources, they inevitably suffer from problems of non-response and sample attrition. As rates of sample non-response and sample attrition increase, this tends to reduce 'external validity', because respondents become less representative of the target population, while differential rates of sample attrition between programme and control groups may adversely affect 'internal validity' if those who drop out of the study in each group have different characteristics.

A number of steps were taken in the design of the ERA Demonstration to limit the impact of low survey response, and these are documented in the next subsection. The focus in this subsection is on the sample size needed to detect statistically significant programme impacts.

Surveys are constrained in detecting programme impacts within an experimental framework for two reasons. The first constraint is the size of the target groups in the programme sites. In the case of the ERA Demonstration, for example, the expected flow into the programme must be capable of yielding a survey sample large enough to allow statistically significant programme impacts to be estimated for each of the three ERA target groups.

The second constraint is that of cost, which clearly limits the number of interviews that can be conducted. Given that the unit cost of each survey interview in the ERA evaluation is expected to be quite high, it will not be possible to conduct more than around 5,000 interviews during each of the two planned post-random-allocation survey waves. The sample of those who will be interviewed will be drawn randomly across the six experimental sites from individuals assigned to the programme and control groups.

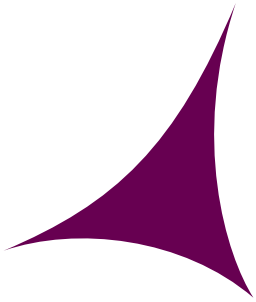
Equations of the sort that appear below are commonly used during the planning phase of a social experiment to help evaluators assess the trade-off between survey cost and sample size:

$$MDI = z\sqrt{\frac{\sigma^2\left(1-R^2\right)}{p(1-p)n}}$$

This equation allows an evaluator to determine a Minimum Detectable Impact (MDI) (Bloom 1995) for a specific outcome (for example, earnings), given estimates of expected sample size and certain statistical assumptions. The MDI is the smallest impact of the evaluated programme that can be reliably estimated. If the MDI exceeds the actual impact of the programme, it will not be possible to determine the size of the programme's impacts with confidence. We discuss the equation in some detail next because it illustrates some of the key issues in experimental design.

As implied by the equation, the greater the sample size, which is represented by '*n*' in the equation, the smaller the MDI will be. In the case of the ERA Demonstration, around 1,600 individuals in each target group, across the six experimental sites, are expected to respond to the 24-month follow-up survey. The quantity '*p*' in the equation represents the proportion of the sample allocated to the programme group, which, in the case of the ERA Demonstration, will be approximately '0.5'. A 50:50 allocation ratio, *ceteris paribus*, produces the smallest MDI, with a higher MDI resulting from ratios either larger or smaller than this.

The estimated population variance of the outcome for which the evaluator is calculating an MDI is represented in the



equation by 'σ^2'. The variance of an outcome is a statistical measure of the extent to which the outcome varies among individuals in a programme's target population. As the equation implies, a larger variance results in a larger MDI. The reason for this is that a different mean value for the outcome is obtained each time a sample is drawn from a population group because each sample will contain a different set of individuals. If the variance of the outcome within the population group is large, these means will vary widely. Impacts calculated on the basis of outcomes with mean values that vary widely from sample to sample are intrinsically harder to detect than those with smaller variances, and therefore require larger samples.

The obvious problem facing evaluators, who are using the equation to estimate the anticipated MDIs for different programme impact measures, is how to obtain a value for 'σ^2' prior to having the actual experimental data. In order to get a reasonable estimate of the variance, evaluators need to consult previous studies where data have been collected on similar outcomes. These studies can be previous evaluations or large probability surveys (Orr 1999). Variance estimates should be obtained from samples comprising individuals very similar to the target groups for the experiment and measured over a similar time period. For example, if an impact is to be the average difference of some continuous outcome, such as earnings over a 12-month period, then the estimated variance of the outcome should be calculated for a 12-month average.

The quantity '*z*' in the equation represents a multiplier that converts the estimated standard error for an outcome into an MDI. It is the sum of the '*z*'-values, drawn from a standard normal cumulative distribution of mean zero, for the required statistical significance and statistical power of the test used to measure the impact. Many social experiments assume 95 per cent statistical significance and 80 per cent power. For a one-tailed statistical test, this would equate to a value for '*z*' of 2.49 (that is, 1.65 for statistical significance plus 0.84 for statistical power).

It is important that evaluators are able to justify chosen levels of statistical significance and power. For example, in specifying 95 per cent statistical significance, the evaluator is accepting a 5 per cent chance of erroneously rejecting the null hypothesis of no impact, when in fact, a true impact exists. This is known as Type 1 statistical error. Similarly, specifying 80 per cent power assumes a willingness to accept a 20 per cent chance of failing to reject the null hypothesis, when in fact the programme has had an impact – a Type 2 statistical error. The trade-off between these two types of error is important because the costs associated with each are different. As Orr (1999) points out, the costs associated with introducing a programme that does not work (implied by making a Type 1 error) are greater than the costs of not introducing a programme that does work (implied by making a Type 2 error). With a Type 1 error, the full costs of the programme are incurred, but there are no benefits. With a Type 2 error, in contrast, programme benefits are lost, but no programme costs are incurred. The ERA Demonstration explicitly requires statistical significance of 95 per cent, and power of 80 per cent, implying that the risk of a Type 2 error is considered less costly than the risk of a Type 1 error by a factor of 4:1.

Evaluators often recommend two-sided statistical tests when estimating impacts, on the assumption that a programme could

have either a positive or negative effect. Some evaluators, such as Bloom (1995), however, argue that a one-sided statistical test is more appropriate, because unlike research that aims to estimate a relationship between two variables, social experiments are conducted to determine whether the tested programme has produced the impacts that were intended. This implies that the null hypothesis for many social experiments should be no impact and, depending on the objectives of the tested programme, the alternative hypothesis either a positive or negative impact. Because the objectives of ERA are to increase employment, job retention, and job advancement, MDIs for these outcomes were estimated, as Bloom suggests, on the basis of a one-tailed statistical test. The adoption of a one-tailed test, *ceteris paribus*, results in a smaller MDI than when a two-tailed test is assumed.

Most social experiments measure programme impacts through a simple comparison of means or proportions with a corresponding t-test or chi-squared test for statistical significance, or through the use of a linear regression model, similar to the one set out below:

$$Y_i = \beta_0 + \beta_1 P_i + \sum_k \beta_{2k} X_{ik} + \varepsilon_i$$

In this equation, 'Y_i' represents a continuous variable measuring an outcome (for example, earnings) for the 'ith' unit. 'P_i' is a binary indicator variable that is given a value of 1 when the unit is allocated to the programme group and zero otherwise. 'X_{ik}' represents a set of 'k' independent variables measured prior to random allocation, at baseline, for each unit and hypothesised on the basis of theory or evidence to affect 'Y'. The estimated coefficient 'β_1' represents the impact of the programme on 'Y' and possesses the same expected value as the simple average difference between mean outcomes in the programme and control groups. The reason for using a regression model to estimate programme impacts is that statistical precision is improved because residual differences between the programme and control groups in the 'X's, which remain despite random allocation, are controlled for by the regression.

In the MDI calculations for the ERA Demonstration project, the use of regression adjustments in estimating impacts was accounted for by including the term '$(1–R^2)$' in the first equation above. 'R^2' represents the explanatory power of the linear regression (the proportion of the variance in 'Y_i' explained by the regressors); a larger 'R^2' results in a smaller MDI.

Obtaining data on outcome measures

As previously mentioned, both experimental and non-experimental evaluations in the UK are heavily reliant on survey data in order to measure a full range of programme outcomes. Data on certain important outcomes, such as wages and earnings, are rarely available from administrative sources. In the UK. It is also difficult to obtain accurate information on the destinations of individuals leaving the benefits system from administrative records. In theory, administrative data are unaffected by non-response and sample attrition. In practice, however, administrative data sets are subject to missing-data problems, but usually not to the same extent as surveys.

Compared to North America, evaluation research in the UK has suffered from relatively poor levels of survey non-response. For example, in North America, response rates of 80 per cent are not uncommon for five-year follow-up surveys (Morris *et al*. 2003). In the UK, however, major evaluations, such as the ONE evaluation, recorded survey response rates of 73 per cent for a baseline survey (Green *et al*. 2000) and only 59 per cent[2] for a follow-up survey conducted approximately six months later (Green *et al*. 2001). An example of a social experiment in the UK is the Restart evaluation. Restart had an initial sample of some 8,000 individuals (White and Lakey 1992), but only 3,400 (42 per cent) of these individuals responded to both of the two follow-up surveys, which were conducted about six months apart.

Poor survey response threatens the validity of results from social experiments. As a result of initial non-response and sample attrition, the achieved sample can be systematically different in terms of the characteristics of its sample members compared to the characteristics of members of the population from which it was drawn. When this occurs, the external validity of impact estimates is called into question. Moreover, when the processes of initial survey non-response and sample attrition differ between programme and control groups, selection bias can be re-introduced into the data, undermining the experimental design and calling into question the internal validity of impact estimates.

When survey data suffer from the problems outlined above, survey non-response weights and quasi-experimental econometric methods can be used to attempt to recover unbiased programme impact estimates at the data analysis stage. However, it is always preferable to build mechanisms into the data collection design that minimise survey non-response and sample attrition before they arise. We next outline the ERA Demonstration project's approach to survey data collection, discussing measures that are being adopted to ameliorate the problems of non-response and sample attrition. The measures described below apply equally to programme and control group and, taken together, represent a concerted effort to maximise response and minimise sample attrition.

The ERA Demonstration will administer a baseline survey immediately prior to the point at which individuals are randomly assigned. In addition, the design calls for surveys to take place 12 and 24 months after random allocation, and possibly, if rates of survey response are deemed likely to be maintained, at 60 months after random allocation, at which times data on outcomes will be collected. The initial evaluation design specified that individuals participating in ERA Demonstration surveys will be paid for completing each of the follow-up survey interviews.

For the purposes of the baseline survey, each individual will complete a questionnaire referred to as the Baseline Information Form or BIF. As part of the baseline survey, each individual will be asked to consent to being asked to take part in ERA research. They will also be asked to give their consent to being randomly assigned and taking part in the ERA project. Those individuals who refuse to be randomly assigned or consent to participation in follow-up surveys or fail to complete the BIF will not be allowed to enter the study. Consequently, baseline measures in the form of data from the BIF will be available for all individuals who are randomly assigned. However, if a large fraction of individuals eligible for the programme refuse to

[2] The response rates calculated on the basis of the eligible population at wave 1 would be lower than 59 per cent.

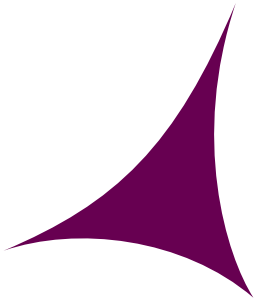
complete the BIF or to give their consent to take part in the research, or to be randomly assigned, this will have a deleterious effect on the external validity of the ERA evaluation. As a result, programme staff will be encouraged to sell the benefits of participating in the study as strongly as possible. These benefits include the chance to be allocated to the programme group and, as a consequence, to receive new ERA services (including financial incentive payments); the compensation that individuals will receive for the time they spend participating in survey interviews; and the fact that individuals have an opportunity to contribute information that will be used to plan services that affect them and their peers.

The ERA Demonstration's approach to survey data collection incorporates further steps to improve contact rates, as well as mechanisms for reducing rates of refusal. Tracing individuals in order to interview them is a challenge for survey research, especially for longitudinal surveys among low-income groups. Many evaluations of welfare-to-work programmes in the UK have relied on samples drawn from benefit records. Benefit records, however, do not always contain up-to-date address and telephone details, a fact that can lead to high rates of non-contact. Thus, the ERA evaluation will attempt to generate an entirely new and up-to-date address and telephone record for each individual entering the study. To attempt to ensure that accurate contact details are available at the point at which individuals are randomly allocated, the address details and postcode given by individuals will be checked electronically to determine whether they match up. In addition, on entry into the programme, individuals will be asked to provide contact details for two or three relatives or friends so that they can then be traced through these relatives or friends, should they move and leave no forwarding address. This information will be entered on each individual's BIF.

Other measures to improve contact rates include having an extended contact window of six months for surveys. This involves allowing sampled individuals to be surveyed for up to six months past the 12- and 24-month anniversaries of their random allocation. In addition, benefit records are to be checked for changes in contact details prior to conducting survey interviews. Such a check can provide an alternative address to visit or telephone number to call if the contact details on an individual's BIF record prove to be out of date or inaccurate. The design also calls for survey interviewers to attempt to contact the entire sample between survey waves, first by telephone and then, if that fails, by making face-to-face contact. At these between-wave contacts, interviewers will ask each respondent whether they have any plans to move, as well as updating the contact details for the respondent's two or three relatives or friends.

Other elements of the ERA Demonstration survey design address the issue of refusal to participate in surveys. In addition to the provision of cash incentives for individuals to participate in survey interviews, an attempt will first be made to interview sampled individuals over the telephone. Experience gained in the evaluation of the ONE programme (Morris *et al*. 2003), which involved survey interviews with a similar population, suggests that many individuals in the ERA target groups prefer to be interviewed over the telephone.

When sampled individuals cannot be contacted by telephone or a telephone interview is not possible for other reasons,



a face-to-face interview will be attempted by interviewers specially trained in 'refusal conversion' techniques.

The need for quasi-experimental comparisons

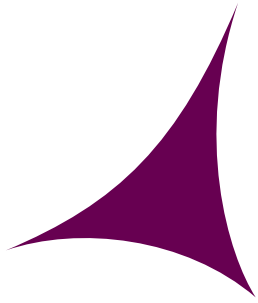
When the impact of certain social phenomena (for example, divorce) need to be evaluated, but the processes giving rise to the phenomenon cannot be controlled or directly manipulated by the evaluator or policymaker, or when social experimentation is otherwise inappropriate, a range of quasi-experimental methods can be implemented. However, even when a social experiment is conducted, this does not necessarily do away with the need for quasi-experimental methods. Indeed, to answer certain questions of interest to policymakers, quasi-experimental methods must be used within an experimental design.

In the case of the ERA project, a random allocation experiment is both feasible and appropriate. Moreover, most of the impacts of interest can be estimated experimentally, that is, by the direct comparison of outcomes for the programme group with outcomes for the control group. However, a few cannot be estimated experimentally. For example, in evaluating ERA, it will be important to determine whether services have a positive impact on wage rates and wage progression. Because wage rates are only available for individuals who work, the programme-group/control-group comparisons of wage rates must be limited to those who have found work. Indeed, examinations of wage progression must rely on individuals in the sample who work, at two separate points in time. Because the ERA treatment may influence who it is that works, the characteristics of those in the programme group with jobs might systematically differ from the characteristics of those in the control group with jobs. If so, the comparison of outcomes between the two groups will not be a randomised comparison; it will instead be a quasi-experimental comparison.

As discussed earlier, the key problem with quasi-experimental comparisons is *selection bias* – that is, the possibility that outcomes differ between the groups being compared because their characteristics differ systematically, rather than because of differences resulting from the treatment being tested. For example, if the ERA treatment helps those nearest the lower margin of employability find and maintain employment, this will reduce the average wage rate of the programme group relative to the average wage rate of the control group because the former will, on average, have characteristics that are less attractive to employers than the latter. This could be due to differences in either 'observables' (i.e. characteristics such as age, race, or education that are readily measured) or 'unobservables' (characteristics such as motivation and self-esteem that are difficult or not feasible to measure).

There are three alternative approaches that might be used to correct for selection bias in making non-experimental comparisons:

1. Assume balancing biases. Here, it is assumed that biases result from restricting the analysis only to members of the programme and control groups who work, because such individuals differ from those who do not work. However, it is further assumed that the biases are similar for both the programme and control groups. Thus, in comparing the working members of the two groups, the biases offset and cancel each other out. Unfortunately, the



'balancing biases' assumption is probably untenable because the ERA treatment means that individuals in the programme group face a different set of circumstances than those in the control group.

2. Assume that there is selection on the observables, but not on the unobservables or, alternatively, that biases resulting from unobservables balance out once the observables are taken into account. If this rather strong assumption holds, it is possible to correct for any differences between working members of the programme and control groups statistically through regression analysis because the sources of the differences (i.e. the observables) between the two groups can be measured.

3. Assume that there is selection on both the observables and the unobservables. In this case, it is necessary to correct for both types of bias. As mentioned under 2., differences between working members of the programme and control group that result from observables can be corrected statistically through regression analysis. It might also be possible to correct for differences between the two groups that result from unobservables by adding a selection term of the sort described by Heckman (1978) to the regression. The selection term itself would be derived from separate probit regression equations in which employment status is regressed against a set of explanatory variables, which differ from the set of explanatory variables included in the wage-rate and wage-progression regressions. The success of this approach depends on how well a set of fairly strong assumptions is satisfied.

As the above discussion suggests, it is somewhat problematic as to whether reliable estimates of those impacts of ERA that *must* be estimated quasi-experimentally can be obtained. It is for this reason that experimental comparisons will be relied on to estimate as many of the programme's impacts as possible.

5. WHAT WILL NOT BE LEARNT FROM THE ERA DEMONSTRATION?

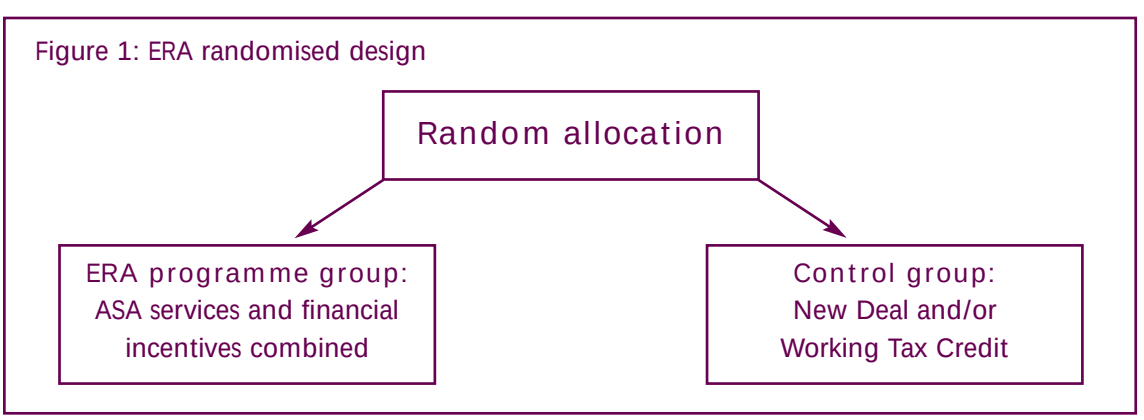
In this section two sets of limitations to social experimentation are discussed: the classic 'black box' problem; and the issue of external validity, or 'generalisability' in social experimentation. In each case, the issues at hand are explored with reference to the design of the ERA Demonstration project.

The black box problem

One of the enduring criticisms of social experiments is that they fail to address the 'black-box' problem (Shadish, Cook and Leviton, 1991; Pawson and Tilley 1997; among others). Social experiments do not provide information as to how the implementation of the programme under consideration affected measured impacts. As part of the evaluation design component of the ERA Demonstration, a full process study is specified which aims to explore the causal mechanisms and the contexts or settings that give rise to the effects measured through the impact study. A problem remains, however. Many welfare-to-work programmes, the effects of which are typically measured through either experimental or quasi-experimental impact studies, comprise a combination of distinct services, delivered as a package. Policymakers often want to know which elements of the package were most effective. In order to address such a question in a rigorous and reliable manner, a more complex experimental design is required.

Figure 1 depicts the ERA experimental design. Individuals entering the ND25+ or the NDLP are randomly assigned to either the ERA programme group or to a control group. Those in the programme group start to receive ASA services prior to entering work. After starting a job, they continue to receive ERA services comprising financial incentives, where they qualify for these, combined with support from an ASA. Members of the programme group can also claim tax credits. Individuals assigned to the

Figure 1: ERA randomised design

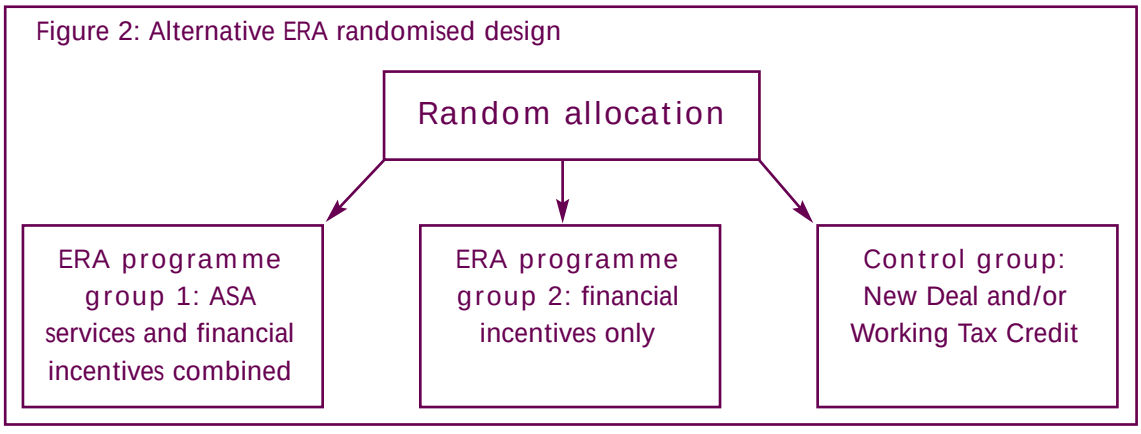


control group enter the New Deal, as they would have if the demonstration had not been in operation. They can also qualify for tax credits on entering work. Lone parents on WTC who enter the demonstration and are assigned to the programme group, continue to qualify for tax credits, and in addition can receive in-work support from an ASA as well as qualifying for ERA financial incentives. Those assigned to the control group continue to receive tax credits.

As discussed previously, comparing the difference between average outcomes in the programme group with those in the control group provides an unbiased estimate of the impact of the combined ERA service. However, such a research design does not indicate whether ERA services were delivered in a manner consistent with the programme's design. Thus, the process study addresses this question. Moreover, this design does not allow the separate contributions of ERA financial incentives and the ASA services to the overall impacts of the programme to be estimated.

Figure 2 illustrates an alternative experimental design that was considered for the evaluation of the ERA Demonstration. This design allows the evaluation to determine the impact of combined ASA services and financial incentives, as well as the impact of financial incentives in isolation. Individuals are randomly allocated to one of three groups. Comparing average outcomes for Programme Group 1 with those in the control group provides an unbiased estimate of the impact of combined ASA and financial incentives. Conversely, an unbiased estimate of the impact of financial incentives alone can be obtained by comparing average outcomes for individuals allocated to Programme Group 2 with those in the control group. Information from an experiment of this kind provides policymakers with the ability to compare the effectiveness of different programme components and, in the case of the ERA Demonstration, would determine whether it is more effective to combine caseworker services with financial incentives or simply introduce financial incentives alone. Impact estimates from such a design, combined with estimates of the net cost of each programme package, can then be used to determine which combination of services, if any, provides the most cost-effective approach to improving retention and advancement. Such a design is often referred to as a 'differential' experimental design.

Figure 2: Alternative ERA randomised design



There are, however, both practical and analytical barriers to implementing a differential experimental design. As previously discussed, experimental designs of all kinds are prone to contamination and crossovers. When complex, differential, experimental designs are used, such as the one illustrated in Figure 2, the potential for crossovers and contamination is increased.

Social experiments make administrative demands on those implementing the programme to be evaluated. Increasing the number of programme groups may increase the probability of administrative confusion and of an individual receiving the wrong set of services. In contexts such as the UK, where there is limited experience in implementing and managing random allocation designs, it may be preferable to keep the design as simple as possible. In North America, where there is a longer history of using random allocation to evaluate social programmes, a number of examples exist where differential designs have been used effectively.

In addition to the practical barriers to implementing a differential design, there is also the problem of sample size to consider. In order to be able to make statistically significant comparisons between outcomes in two programme groups and a control group, all things being equal, the required sample will need to be 50 per cent larger than that necessary for a single-programme-group design. In some cases, the additional sample size required may render a differential design impracticable. In the case of the ERA Demonstration, the number of experimental sites would need to be expanded beyond the six currently planned and it would be very difficult to estimate site-specific impacts.

Evidence from Canada (Michalopoulos *et al*. 2002) as well as the United States (Knox, Miller and Gennetian 2000) suggests that combining financial incentives with caseworker services is likely to produce bigger impacts than financial incentives alone for low-income groups. Thus, given the practical considerations, sample size constraints, and the fact that existing evidence suggests that individuals are likely to benefit more from a package that combines caseworker services with financial incentives, the ERA impact study was designed as a simple, single-programme-group design.

Of course, adopting a single-programme-group design, as set out in Figure 1, limits what can be learned about the individual effectiveness of different ERA programme components. The process study will, however, help in understanding the processes by which the various programme components bring about observed effects.

Generalisability[3]

As previously indicated, a critical issue in the evaluation of Government programmes is 'external validity' – the extent to which estimated programme effects can be generalised to different locations and populations, to different time periods, and to different variants of the programme being studied. Questions about external validity apply almost equally to experimental evaluations, such as that of ERA, and to quasi-experimental evaluations. The external validity of specific estimates of programme effects may be questioned for a number of reasons. One of these, the representativeness of pilot sites, was discussed above. A number of others are considered in this subsection.

[3] Parts of this section borrow from Friedlander, Greenberg and Robins 1997.

Extrapolation to different times and places

This is a serious, if obvious, problem. Social attitudes, Government institutions, the business cycle, the relative demand for unskilled and skilled labour, and other relevant factors may change in the years following an evaluation. Likewise, different locations may have dissimilar social attitudes, local government institutions, labour market conditions, and so forth. Moreover, the characteristics of programme participants could differ as well.

Scale bias

The external validity of pilot tests of policy innovations may be compromised by 'scale bias'. Manski and Garfinkel (1992) and Garfinkel, Manski and Michalopoulos (1992) suggest that when pilot tests are scaled up to universal participation, this could change community norms or combine with patterns of social interaction or information diffusion in ways that will feed back and influence the success of the policy innovation. These community or 'macro' effects, they argue, will be absent in small-scale pilot programmes or partially-scaled programmes. In addition, testing a programme on a small scale may cause the composition of the programme participants to differ from what it would be if the programme were rolled out nationally by inhibiting the diffusion of information about the programme to potential applicants or, in an experiment such as ERA, by discouraging risk-averse individuals from applying to a programme when they could be randomly assigned to a no-services control group (see Heckman 1992; Heckman and Smith 1995; and Manski 1993, 1995). At present, little is known about the practical importance of these effects. Although the possibility of bias caused by distortion of the participant sample in small-scale pilot tests has strong theoretical appeal, its empirical importance is yet to be demonstrated. This issue is further discussed below in considering 'entry effects'.

One quasi-experimental approach for avoiding biases caused by testing policy innovations on a small scale is to implement them on a site-wide, fully-scaled basis in some locations and, for comparison, use other sites (perhaps statistically matched) that have not adopted the innovation. Although this 'saturation' evaluation design does, in principle, allow feedback effects to be captured, the programme may have to be kept in place for many years, with firm guarantees of permanency, before these effects reach full potency. Moreover, as previously discussed, cross-site-comparison designs will produce unreliable estimates of programme effects if the programme and comparison sites differ in ways that are inadequately controlled for in the evaluation.

Services received by control group members

It is often the case that some members of control or comparison groups receive services similar to those received by programme group members. For example, in the case of the ERA evaluation, members of New Deal target groups will receive help in securing employment regardless of whether they are assigned to the programme or control group, although the nature of this help will differ in some respects. Under these circumstances, estimates of programme impacts do not measure the pure effect of participating in the evaluated programme versus the absence of receiving any similar services at all. Rather, they measure the incremental effect of whatever additional services the

programme provides. For example, the ERA programme group will receive two years of post-employment casework services, as well as financial incentives that encourage full-time stable employment and participation in training while working; but the control group will not.

The fact that the services received by the programme and control groups overlap to some degree does not distort programme evaluation findings, as long as the services received by the latter are representative of the true counterfactual. If they are, the resulting impact estimates will clearly be policy-relevant. However, the overlap is a source of at least two potential threats to external validity. First, not only will the evaluated programme differ over time or from one place to another, but the array of activities available to comparison-group members will also differ, complicating the problem of generalising the evaluation results. Second, the very existence of the programme being evaluated might change the services available to the control group. This second threat to external validity, which Heckman and Smith (1995) call 'substitution bias', could occur, for example, if ERA absorbs resources that would otherwise be available to members of the control group or, alternatively, if, as a result of serving some persons who would otherwise enter the New Deal, ERA frees up additional resources that can then be used to serve those who enter the New Deal and are assigned to the control group.

Hawthorne effects

The behaviour of participants in a pilot test of programme or policy could be influenced by knowledge that they are part of the pilot test, not only by the receipt of the tested services, a so-called 'Hawthorne effect'. For example, if ERA participants know that their labour market performance will be measured in terms of certain outcomes, such as stable work patterns, some of them might attempt to succeed in terms of these outcomes.

There is virtually no information about whether Hawthorne effects bias findings from social experiments. It seems possible that members of both the programme and control groups could respond similarly to being part of a social experiment. If so, such effects will cancel out in measuring impacts, and there would be no bias. Alternatively, some control group members could be discouraged by the fact that they were allocated to the control group, rather than the programme group, and alter their behaviour for that reason.

Entry effects

If the services provided by a programme are perceived as beneficial, then some individuals who are initially ineligible to participate may adopt behaviours needed to qualify (an 'entry' effect). On the other hand, in the case of mandatory work or training requirements for benefit recipients, individuals might leave the benefit rolls when they are informed that they will be subject to the newly-established requirements (an 'exit' effect). Similarly, some individuals who might otherwise have entered the benefit rolls may decide not to do so if they will be required to meet work or training requirements (a 'deterrent' effect).

Manski and Garfinkel (1992) and Moffitt (1992, 1996), among others, have argued that programme entry, exit and deterrent effects could be substantial. However, findings from non-experimental attempts to measure these effects, which have generally relied on aggregate-level time-series studies of programme applications, are mixed and inconclusive (for example, see Johnson, Klepinger and Dong 1990; Wissoker and

Watts 1994; Chang 1996; Phillips 1993; Schiller and Brasher 1993). There has been only one attempt to use experimental methods to measure entry effects – an evaluation of a pilot test of a Canadian programme that provided very generous earnings supplements to lone parents on welfare who worked full-time (Berlin *et al*. 1998). Newly enrolled benefit recipients, who were allocated at random to a programme group, were told that if they remained on welfare for the next 12 months, they would subsequently qualify for earnings supplements provided they then worked full-time. The control group was not given this information, as they were not eligible for the earnings supplement. After a year, 3.1 per cent more of the programme group than the control group were still on the welfare roll.

If rolled out nationally, the ERA programme could potentially cause important entry effects among members of each of the three programme target groups. First, while individuals must participate in ND25+ after they have been on Jobseeker's Allowance for 18 months, they can volunteer before then. Although not many individuals in receipt of Jobseeker's Allowance currently volunteer, this may change if the opportunity exists to qualify for the financial incentive payments provided by ERA. Second, the NDLP is a voluntary programme for lone parents who are either not working or working fewer than 16 hours a week. The financial incentives offered by ERA could induce more such individuals to volunteer. Third, lone parents who work part-time (between 16 and 30 hours a week) will be able to qualify for ERA incentive payments, but those working full-time (over 30 hours) will not. Thus, there will be incentives for lone parents who are currently working full-time to temporarily reduce their hours in order to qualify.

None of these entry effects are likely to be important in the pilot test of ERA. Because it will be run in only six sites and enrolment into the pilot test will be limited to a year in most cases, relatively few of those who do not already qualify for the test programme will be sufficiently knowledgeable about it to change their behaviour accordingly. However, this would no longer be the case if the programme were rolled out nationally on a permanent basis. Thus, if entry effects are important, findings from the pilot test may not generalise to a permanent programme. However, a national rollout of ERA might well be accompanied by rules that are specifically designed to minimise entry effects. For example, a national ERA could be restricted to unemployed persons who have been receiving Jobseeker's Allowance for at least 18 months. A similar restriction could be imposed on WTC lone parents who have been working part-time. Of course, some unemployed persons and part-time workers who desire full-time work might wait for 18 months before taking such jobs. However, the evidence mentioned above for the Canadian programme suggests that this effect is likely to be small. If rules that succeed in limiting entry effects were made part of a national ERA, findings from the pilot test are likely to be more generalisable to the permanent programme.

General equilibrium effects

A Government programme that is being pilot tested may have important effects on the wellbeing of those who are not enrolled in the programme, or at least it would were the programme rolled out nationally. Two such effects are equilibrium wage effects and substitution effects. Empirical evidence about the magnitude of both of these effects is quite limited.

If participants in a programme search harder for jobs, or work more weeks or hours than they otherwise would, the resulting increase in labour supply will tend to lower the equilibrium wage within the labour markets in which they work. Thus, workers who are employed in the same labour markets as programme participants could receive lower wages than otherwise. For this effect to be very large, however, three conditions must hold: (1) the minimum wage must not constrain downward movements in wage rates; (2) programme participants must account for a fairly large share of the workers in the relevant labour markets; and (3) programme effects on job search and weeks and hours worked must be fairly large.

Even if rolled out nationally, ERA seems unlikely to bring substantial equilibrium wage effects. It is anticipated that, at least initially, most participants would be employed in low-wage labour markets. Thus, at least to some degree, the minimum wage would probably constrain reductions in equilibrium wages. Moreover, the ERA target groups are limited to the long-term unemployed and lone parents. Because the long-term unemployed participate in the same labour markets as other unemployed persons and individuals who are currently employed, and lone parents participate in the same labour markets as married and childless persons, they account for only a fairly small proportion of the total supply population in any given labour market. Finally, ERA's impacts are expected to be moderate at best.

Substitution effects occur if participants in a programme hold jobs that individuals who do not participate would otherwise have held (Johnson 1979). If these non-participants become unemployed or accept lower-wage jobs as a result, then their earnings fall.

Despite these potential adverse effects, there is very little research quantifying the magnitude of substitution effects. However, a recent evaluation of the New Deal for Young People (NDYP) provides a preliminary analysis of substitution effects that suggests they could be modest (Blundell *et al*. 2002).

In the case of ERA, substitution effects would occur if the intervention has a positive impact on the job retention or job advancement of those who are included in the target group, and, as a result, fewer job vacancies or opportunities for advancement were available to those who are not included in the ERA target group. The magnitude of this potential substitution effect is likely to depend on the state of the local labour markets in the programme pilot sites. If a local labour market is tight, then alternative job opportunities are likely to be available to those outside the target group; but if it is loose, then the cost of substitution to those affected could be substantial.

It is also possible that, as a result of its emphasis on advancement, ERA will help some participants to leave slack occupational labour markets for tight ones – for example, through encouraging training. If this occurs, ERA would decrease the competition for job vacancies in the slack markets, making it easier for those who remain in these markets to find jobs. In theory at least, this could produce a result that is the exact opposite of a substitution effect: total employment among those not participating in ERA could actually increase

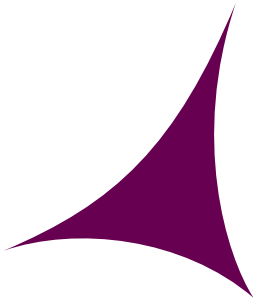
6. CONCLUSIONS

Using the planned evaluation of the ERA Demonstration for illustrative purposes, this paper examines the strengths and weaknesses of random allocation experiments for evaluations of social programmes. The ERA Demonstration will test the efficacy of a new policy intervention that combines pre-employment and in-work support with financial incentives to attempt to find jobs for those who need them and to sustain employment and facilitate advancement in jobs for those who are in employment. These services and financial incentives will be tested in six Jobcentre Plus districts, which differ from one another in a variety of ways and are located throughout Great Britain. The services and financial incentives are targeted at three groups of disadvantaged individuals: participants in the ND25+, participants in the NDLP, and WTC claimants.

The paper suggests that, for evaluating ERA and a wide variety of other social policy interventions, an experimental design is superior to alternative designs that might be used instead – for example, 'before and after', 'matched sites', or 'participant/non-participant' comparisons. It will provide greater assurance of internal validity, while being no more costly or time-consuming. However, this does not mean that an experimental design is always superior for evaluating all social policies; just that it is often advantageous and that it is clearly so for evaluating ERA. Non-experimental methods may be less expensive and less time-consuming, however, than random allocation for evaluating already existing programmes. Moreover, occasionally there are ethical reasons for not using random allocation. Nonetheless, if implemented and run properly, an experimental design will almost always provide greater internal validity.

No single evaluation design, even random allocation, can answer all the questions about a specific social policy that are of interest. Sometimes, however, certain design modifications can be made that can help address certain issues. For example, although ultimately not adopted, consideration was given to using a differential experimental design for the ERA Demonstration in order to determine whether the impact of combining financial incentives with services would be greater than the impact of financial incentives alone. Other limitations of a single evaluation design can be at least partially overcome by combining several different approaches. As discussed in the paper, for example, non-experimental econometric methods will be required to examine certain issues concerning ERA's impact on advancement, while a process analysis will be used to help determine whether ERA services were delivered in the manner intended.

There are certain important questions that no combination of evaluation methods can definitively address, however. As detailed in the paper, for example, neither experimental



nor non-experimental methods will be able to provide more than limited information about which specific components of ERA are most or least effective – the so-called 'black box problem'. In addition, once findings from the ERA Demonstration become available, uncertainty will inevitably remain about their external validity – that is, the extent to which they can be generalised to different locations and populations and to different time periods; whether they are subject to scale bias, general equilibrium wage effects, substitution effects, and/or Hawthorne effects; and whether entry effects might occur if ERA is rolled out nationally that did not arise during the Demonstration – regardless of the combination of experimental and non-experimental methods that were used to obtain them.

7. REFERENCES

Arulampalam, W. and Booth, A. (1998) 'Training and Labour Market Flexibility: Is there a Trade-off?', Institute for Labour Research Working Paper No.13 (Colchester: University of Essex).

Berlin, G., Bancroft, W., Card, D., Lin, W. and Robins, P.K. (1998) *Do Work Incentives have Unintended Consequences? Measuring 'Entry Effects' in the Self-Sufficiency Project* (Ottawa: Social Research Demonstration Corporation).

Bloom, H.S. (1995) 'Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs', *Evaluation Review*, 19, 547–56.

Bloom, H.S., Michalopoulos, C., Hill, C. and Lei, Y. (2002) *Can Non-experimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?* (New York: Manpower Demonstration Research Corporation).

Blundell, R., Costa Dias, M., Meghir, C. and Van Reenen, J. (2002) 'Evaluating the Employment Impact of a Mandatory Job Search Program: The New Deal for Young People in the UK', unpublished manuscript (London: University College London and Institute for Fiscal Studies).

Boruch, R.F. (1997) *Randomized Experiments for Planning and Evaluation: A Practical Guide* (Thousand Oaks: Sage Publications).

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

Burtless, G. and Orr, L.L. (1986) 'Are Classical Experiments Needed for Manpower Policy?', *The Journal of Human Resources*, 21, 607–39.

Campbell, D. and Green, F. (2002) 'The Long-term Pay-off from Working Longer Hours', Department of Economics Discussion Paper 0205 (Canterbury: University of Kent).

Campbell, D.T. and Stanley, J.C. (1963) *Experimental and Quasi-experimental Designs for Research* (Boston: Houghton Mifflin Company).

Chang, F. (1996) *Evaluating the Impact of Mandatory Work Programs on Two-Parent Welfare Caseloads*, Doctoral dissertation, (Baltimore: University of Maryland, Baltimore County).

Cook, T.D. and Campbell, D.T. (1979) *Quasi-experimentation*: *Design and Analysis Issues for Field Settings* (Boston: Houghton Mifflin Company).

Fraker, T. and Maynard, R. (1987) 'The Adequacy of Comparison Group Designs for Evaluations of Employment Related Programs', *Journal of Human Resources*, 22, 194–227.

Friedlander, D., Greenberg, D.H. and Robins, P.K. (1997) 'Evaluating Government Training Programs for the Economically Disadvantaged', *Journal of Economic Literature*, December, 1809–55.

Friedlander, D. and Robins, P.K. (1995) 'Evaluating Program Evaluations: New Evidence on Commonly used Non-experimental Methods', *American Economic Review*, 85, 923–37.

Garfinkel, I., Manski, C.F. and Michalopoulos, C. (1992) 'Micro Experiments and Macro Effects', in C.F. Manski and I. Garfinkel (eds) *Evaluating Welfare and Training Programs* (Cambridge, MA: Harvard University Press), 253–76.

Glazerman, S., Levy, D.M. and Myers, D. (2002) 'Nonexperimental Replications of Social Experiments: A Systematic Review', paper presented at the *Annual Meeting of the Association for Public Policy Analysis and Management*, 8 November, Dallas, Texas.

Green, H., Connolly, H., Marsh, A. and Bryson, A. (2001) *The Medium-term Effects of Voluntary Participation in ONE*, Department for Work and Pensions Research Report No. 149 (Leeds: Corporate Document Services).

Green, H., Smith, A., Lilly, R., Marsh, A., Johnson, C. and Fielding, S. (2000) *First Effects of ONE*, Department for Work and Pensions Research Report No. 126 (Leeds: Corporate Document Services).

Greenberg, D.H. and Shroder, M. (1997) *The Digest of Social Experiments* (Washington: Urban Institute).

Heckman, J.J. (1978) 'Dummy Endogenous Variables in a Simultaneous Equation System', *Econometrica*, 46, 931-59.

Heckman, J.J. (1992) 'Randomization and Social Policy Evaluation', in C. F. Manski and I. Garfinkel (eds) *Evaluating Welfare and Training Programs* (Cambridge, MA: Harvard University Press), 201–30.

Heckman, J.J. and Smith, J.A. (1995) 'Assessing the Case for Social Experiments', *Journal of Economic Perspectives*, 9, 85–110.

Heckman, J.J., Ichimura, H. and Todd, P.E. (1997) 'Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme', *Review of Economic Studies*, 64, 605–54.

Hollister, R.G. and Hill, J. (1995) 'Problems in the Evaluation of Community-Wide Initiatives', in J. P. Connell, A. C. Kubisch, L. B. Schorr and C. H. Weiss (eds) *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts* (Washington DC: Aspen Institute), 127–72.

Johnson, G. (1979) 'The Labor Market Displacement Effects in the Analysis of the Net Impact of Manpower Training Programs', *Research in Labor Economics*, Supplement 1, 227–54.

Johnson, T.R., Klepinger, D.H. and Dong, F.B. (1990) 'Preliminary Evidence from the Oregon Welfare Reform Demonstration', unpublished paper, June.

Knox, V., Miller, C. and Gennetian, L. (2000) *Reforming Welfare and Rewarding Work: Final Report of the Minnesota Family Investment Program* (New York: Manpower Demonstration Research Corporation).

LaLonde, R.J. (1986) 'Evaluating the Econometric Evaluations of Training Programs with Experimental Data', *American Economic Review*, 76, 604–20.

LaLonde, R.J. and Maynard, R. (1987) 'How Precise are Evaluations of Employment and Training Programs? Evidence from a Field Experiment', *Evaluation Review*, 11, 428–51.

Manski, C.F. (1993) *What do Controlled Experiments Reveal about Outcomes when Treatments Vary?,* Institute for Research on Poverty Discussion Paper No. 1005–93 (Madison: University of Wisconsin).

Manski, C.F. (1995) *Learning about Social Programs from Experiments with Random Assignment of Treatments*, Institute for Research on Poverty Discussion Paper No. 1061–95, (Madison: University of Wisconsin).

Manski, C.F. and Garfinkel, I. (1992) 'Introduction', in C.F. Manski and I. Garfinkel (eds) *Evaluating Welfare and Training Programs* (Cambridge, MA: Harvard University Press), 1–22.

Michalopoulos, C., Tattrie, D., Miller, C., Robins, P.K., Morris, P., Gyarmati, D., Redcross, C., Foley, K. and Ford, R. (2002) *Making Work Pay: Final Report on the Self-Sufficiency Project for Long-Term Welfare Recipients* (Ottawa: Social Research and Demonstration Corporation).

Moffitt, R.A. (1992) 'Evaluation Methods for Program Entry Effects' in C.F. Manski and I. Garfinkel (eds) *Evaluating Welfare and Training Programs* (Cambridge, MA: Harvard University Press), 231-52.

Moffitt, R.A. (1996) 'The Effect of Employment and Training Programs on Entry and Exit from the Welfare Caseload', *Journal of Policy and Management*, 15, 32-50.

Moffitt, R.A. (2002) 'The Role of Randomized Field Trials in Social Science Research: A Perspective from Evaluations of Reforms of Social Welfare Programs', paper presented at the *Conference on Randomized Experimentation in the Social Sciences*, 20 August, Yale Institute for Social and Policy Studies.

Morris, S., Greenberg, D., Riccio, J., Mittra, B., Green, H., Lissenburgh, S. and Blundell, R. (2003) *Designing a Demonstration Project – An Employment, Retention and Advancement Demonstration for Great Britain*, Government Chief Social Researcher's Office Occasional Papers Series No. 1 (London: Cabinet Office).

Orr, L.L. (1999) *Social Experiments: Evaluating Public Programs with Experimental Methods* (Thousand Oaks: Sage Publications).

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation* (London: Sage Publications).

Phillips, E.H. (1993) *The Effect of Mandatory Work and Training Programs on Welfare Entry: The Case of GAIN in California*, Doctoral dissertation (Madison: University of Wisconsin).

Purdon, S. (2002) *Estimating the Impact of Labour Market Programmes*, Department for Work and Pensions Working Paper No. 3 (London: HMSO).

Rosenbaum, P.R. and Rubin, D. (1984) 'Reducing Bias in Observational Studies using Sub-classification on the Propensity Score', *Journal of the American Statistical* Association, 79, 516–24.

Rossi, P.H., Freeman, H.E. and Lipsey, M.W. (1999) *Evaluation: A Systematic Approach* (Thousand Oaks: Sage Publications).

Schiller, B.R. and Brasher, C.N. (1993) 'Effects of Workfare Saturation on AFDC Caseloads', *Contemporary Policy Issues*, 11, 39–49.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (Boston: Houghton Mifflin Company).

Shadish, W.R., Cook, T.D. and Leviton, L.C. (1991) *Foundations of Program Evaluation: Theories of Practice* (Newbury Park: Sage Publications).

Weiss, C.H. (1998) *Evaluation* (Upper Saddle River: Prentice-Hall).

White, M. and Lakey, J. (1992) *The Restart Effect: Does Active Labour Market Policy Reduce Unemployment?* (London: Policy Studies Institute).

Wissoker, D.A. and Watts, H.W. (1994) *The Impact of FIP on AFDC Caseloads* (Washington: The Urban Institute).

ANNEX – SUMMARIES OF WELFARE-TO-WORK AND EMPLOYMENT POLICY

Social Experiments in the UK

The Benefits Agency Visiting Officer (BAVO) pilot evaluation

In this pilot test, which began in the United Kingdom in the spring of 2000 and continued for six months, Benefits Agency visiting officers made and maintained direct contact with a small sample of unemployed lone parents receiving Income Support. Through a series of face-to-face interviews the visiting officers attempted to assist subjects in returning to work or obtaining training. The pilot test was to be evaluated by random allocation. However, of the 406 individuals who agreed to participate in the demonstration, 189 were randomly assigned to the treatment group and 217 to the control group. In the event, only fractions of these groups were actually interviewed by the evaluators: 101 (53 per cent) from the treatment group and 140 (65 per cent) from the control group. Thus, the planned random allocation evaluation was not feasible.

The Employment Zones evaluation

The evaluation assessed a pilot programme that is operating in 15 high-unemployment areas. The random allocation evaluation, which began in the year 2000, was conducted in four of these areas. Employment Zones are mandatory for long-term Jobseeker's Allowance recipients and operated by private sector contractors, who provide pre-work support and up to 13 weeks of post-work support. In the Employment Zones, a client and a personal advisor first develop an Action Plan, and then the client undertakes the prescribed actions. Employment Zone contractors receive incentive payments for each client placed in employment within 39 weeks and for each placed client who retains employment for at least 13 weeks. A substantial fraction of the control group could not be traced for the purposes of obtaining outcome data. The Government has not yet released a report describing the experimental findings.

The In-Work Training Grant (IWTG) pilot

Begun in June 2000 and continued for 12 months, the pilot tested programme offered grants for training of up to £750 to lone parents participating in the New Deal for Lone Parents programme at the time they moved from unemployment into jobs. The pilot test was to be evaluated by random allocation, but ultimately was not because of the low take-up of the grants.

The Intensive Gateway Trailblazers (IGT) evaluation

Conducted in 1999, the experiment tested a mandatory 2-week intensive course undertaken during the NDYP gateway. Subjects were followed for nine months. There is no available written report on the impact analysis. However, the evaluators indicate that there was 'some indication that

at the margin IGT increased the proportion of young adults entering jobs'. In practice, the services received by clients enrolled in the treatment group did not differ very much from those received by control clients. For example, there was difficulty in securing attendance at the mandatory courses. As a result, programme impacts were expected to be small.

The 1-2-1 for the very long-term unemployed study

Conducted from June 1996 to April 1997, this experiment tested the effects of voluntary interviews, assessment, and guidance and job search on a medium-sized sample of long-term unemployed (30 months or more) workers. Subjects were followed for six months. Six months into the treatment, the likelihood of exit from the unemployment register increased by 13 percentage points; 34 per cent of the treatment group exited the unemployment register versus 21 per cent of the controls. This impact estimate is statistically significant. Most of the exits were into training and education, rather than employment.

The 1-2-1/workwise for 18–24-year-olds tracking study

Conducted from April 1994 to April 1996, this experiment tested the effects of mandatory interviews, assessment, guidance and job search on a medium-sized sample of long-term unemployed youths. Subjects were followed for two years. Between 13 and 24 weeks after random allocations, 35 per cent of clients in the treatment group either found work or training and education; 22 per cent of the controls achieved these objectives. This 13 percentage point difference is statistically significant.

The lone parent caseworker pilots evaluation

Conducted from 1994 to 1995, this experiment tested a pilot unemployment assistance programme on a medium-sized sample of single-parent welfare recipients. Subjects were followed for eight months. Virtually the same proportion of treatment subjects and control subjects did not change their job status and, thus, did not change their benefit receipts. Of those who did change their circumstances, there was no statistically significant difference between treatment subjects and controls concerning why they changed their official status. No employment effects could be attributed to the intervention.

The Jobplan evaluation

Conducted in 1993, this experiment tested the effects of an intensive goal-setting workshop on a large sample of unemployed workers. Subjects were followed for three months. Compared to the control group, 5.2 per cent more of the Jobplan members were off the unemployment register 16 weeks after random allocation. Of this, 2.1 percentage points are attributed to those who found work; the remainder is equally divided between those who enrolled in work training and other reasons. Only the combined effect is statistically significant.

The Supportive Caseloading evaluation

Conducted in 1993, this experiment tested the effects of mandatory consultation sessions on a medium-sized sample of the unemployed. Subjects were followed for up to 26 weeks. At 13 weeks, 40 per cent of the treatment group was no longer receiving unemployment benefits, compared to 23 per cent of the control group. At 26 weeks, 22 per cent of the treatment group

had found employment compared to 8 per cent of the control group, a 14 percentage-point difference, which was 2 percentage points higher than the effect observed at 13 weeks.

The evaluation of the 30-month-plus Restart interviews

Conducted from May 1992 to December 1992, this experiment tested the effects of using more experienced advisers in conducting Restart interviews on a large sample of Unemployment Benefit claimants. Subjects were followed for six months. There were no statistically significant differences in outcomes between the programme group and the control group.

The evaluation of the 13-week review

Conducted in late 1991, this experiment tested a new unemployment insurance programme on a large sample of individuals collecting unemployment benefits. Subjects were followed for six months. During the six-month tracking period, 46 per cent of the claimants in the programme group left the unemployment register, thereby terminating their receipt of unemployment benefits; in comparison, only 41 per cent of the claimants in the control group left the register. Fifteen per cent of the claimants in the control group who left the register later signed on again, compared with 19 per cent in the programme group.

The Restart experiment

Conducted from 1989 to 1991, this experiment tested the effects of mandatory interviews with a counsellor on a large sample of Unemployment Benefit claimants. Subjects were followed for one year after the interview was scheduled. Restart reduced unemployment claims by around 5 per cent. Persons in Restart spent less time as unemployed claimants during the study period and took less time to leave the unemployed claimant register and find employment or enter a training programme. Restart had an effect on time in a training programme, but not on the use of job search or on wage levels, job stability, or job quality. The analysis of whether Restart affected time in employment was inconclusive. Restart tended to move participants into a non-claimant, non-employment status for a short period immediately after the counsellor interviews, but in the longer term this effect was reversed

The Government Chief Social Researcher's Office

The Government Chief Social Researcher's Office (GCSRO) is based in the Prime Minister's Strategy Unit and co-ordinates and promotes social research across government. It encourages departments to commission the right research at the right time in order to promote evidence-based policy making and the effective use of social research. It ensures that government research is of the highest quality and uses the most appropriate and up-to-date methods and techniques. GCSRO helps ensure that the government social research service has access to people with the right skills. The office maintains effective links with other professional groups within government as well as with the academic community and those engaging in applied social policy research and evaluation outside government. Sue Duncan is the Government Chief Social Researcher.

A web version of the research can be found on Policy Hub (http://www.policyhub.gov.uk). Policy Hub is a web resource launched in March 2002 that aims to improve the way public policy is shaped and delivered. It provides many examples of initiatives, projects, tools and case studies that support better policy making and delivery and provides extensive guidance on the role of research and evidence in the evaluation of policy.

Other publications in the GCSRO's Occasional Papers are:

Morris, S., Greenberg, D., Riccio, J., Mittra, B., Green, H., Lissenburgh, S. and Blundell, R. (2003) *Designing a Demonstration Project – An Employment, Retention and Advancement Demonstration for Great Britain*, GCSRO Occasional Papers Series No. 1 London: Cabinet Office.

Atkinson, J. and Williams, H. (2003) *Employer Perspectives on the Recruitment, Retention and Advancement of Low-pay, Low-status Employees,* GCSRO Occasional Papers Series No. 2 London: Cabinet Office.

Strategy Unit, Admiralty Arch, The Mall, London SW1A 2WH

Tel: 020 7276 1881
Email: strategy@cabinet-office.x.gsi.gov.uk
Website: www.strategy.gov.uk