Review of Government Evaluations: A report for the NAO

Steve Gibbons (LSE and SERC), Sandra McNally (University of Surrey and CEP) and Henry Overman (LSE and SERC)

EXECUTIVE SUMMARY

This report assesses the quality of cost-effectiveness evaluations. It forms part of a wider National Audit Office project on the use of cost-effectiveness evidence in Government.

The report is based on retrospective review of a selection of 35 UK government evaluations in the policy areas of active labour markets, business support, education and spatial policy. We were asked to highlight the strengths and weaknesses of these reports, to assess the robustness and the usefulness to policy makers and to suggest improvements (drawing on the review and our wider knowledge of the evaluation literature).

These four policy areas were selected because policies in these areas are targeted differently (e.g. some at firms, some at individuals) which helps illustrate how evaluation deals with a number of crucial methodological issues. These are also areas where external perception of quality varies markedly and where the team has considerable expertise. Within these broad areas we picked specific evaluations based on a number of factors including the scale of the policy and the evaluation and the methodological approach adopted. Our assessment was based only on the published details of the evaluation reports rather than any additional information provided by departments.

We found that the quality of cost-effectiveness reports varies widely both within and across policy areas. We found evidence of high quality evaluations in the areas of active labour markets and education. In contrast, evaluations in the areas of business support and spatial policy were considerably weaker. On the basis of the set of reports that we reviewed, our overall assessment would be that none of the business support or spatial policy evaluations provided convincing evidence of policy impacts. In contrast, 6 out of 9 of the education reports and 6 (arguably 7) out of 10 labour market reports were of sufficient standard to have some confidence in the impacts attributed to policy.

We make a number of recommendations about the evaluation design (including a number of technical recommendations). We suggest that the use of a control group (or a counterfactual) should be considered a necessary (although not sufficient) requirement for robust impact assessment and value for money calculations. In the areas of business support and spatial policy we feel that more use could be made of administrative data (following examples provided by labour markets and education evaluations). Our more technical recommendations cover the issues raised by the problem of selection (e.g. when people opt in to treatment) and the techniques used to correct for this selection; the need to avoid basic mistakes in

implementation; the need for improvements in the handling of inference (i.e. how certain we are about the effects of policy) as well as the interpretation of impact estimates (e.g. do they apply to the population as a whole, or to a subset of the population).

Turning to the question of cost-effectiveness we recommend that value for money calculations should not be presented unless the impact evaluation meets minimum standards. We also highlight the problems caused by the availability of cost data and the tendency of some internal cost-effectiveness assessments not to be published alongside the external evaluation report.

We also make recommendations concerning the presentation of evaluation evidence, which are aimed at bringing weaker reports up to the standard of the higher quality reports. In particular, we suggest that a technical appendix, written for a specialist audience, should be a core component of every impact evaluation. We also feel that more care should be taken to distinguish between the analysis of programme delivery and the assessment of impact and value for money.

We consider the options available when robust impact evaluation is not possible, raising the possibility that a broad ranging evaluation may not represent good value for money. Possibilities include better use of process evaluation and monitoring or narrowing the scope of evaluations to focus on specific policy features and the impact on specific groups of recipients.

The widespread use of these options may be undesirable because they will reduce the availability of wide-ranging cost-effectiveness evidence. There is no 'magic bullet' solution to this problem, but greater recognition of the need to consider evaluation issues at the time of policy design would help. Similarly, the overall quality of evaluation could be improved by greater use of independent bodies responsible for peer review or more long term evaluation of major policy initiatives (which would be facilitated by establishing better protocols for the confidential sharing of administrative data with trusted researchers).

Overall, our review found a number of very good evaluations that already satisfy many of these requirements. Many more of the evaluations we considered could be easily improved by implementing a number of these recommendations while keeping the same evaluation design. For other reports, particularly in the area of business support and spatial policy, the research design itself is quite problematic and more careful analysis or write-up within that design are unlikely to improve the overall robustness of the evidence. We recognise that for some policies evaluation is, arguably, more difficult, but the gulf between best practice and the evaluations cannot be attributed to this alone. Indeed, in some situations, the structure of the programme and the data collected for the evaluation would have allowed for careful impact evaluation, but this did not happen. In other situations, use of available administrative data and better methodologies could have provided far more convincing evidence.

The issues this report raises need urgently addressing if we wish to produce cost-effectiveness evidence that is fit for purpose.

1. Context

The National Audit Office (NAO) asked us to undertake a study to assess the quality of cost-effectiveness evaluation as part of a wider project on the use of cost-effectiveness evidence in Government.

The main objectives of this work were:

- To identify a list of approximately 40 existing (cost-effectiveness) evaluations, around 10 evaluations in each of 4 policy areas, for retrospective review.
- To highlight the strengths and weaknesses of these reports and to assess the robustness of these studies, specifically concerning their usefulness for policy makers in informing value for money assessments.
- To identify missed opportunities, to suggest improvements and to outline the likely costs and benefits of those improvements (drawing on the review and wider knowledge of the evaluation literature including, where appropriate, comparisons to evaluations commissioned by overseas organisations)

2. Approach

Early on in the process we met with NAO and agreed a template that would be used to structure our retrospective reviews. We were agreed that, in addition to providing basic information about each evaluation (policy objectives, overall scope and methodology) the reviews would (1) assess the extent to which insights from the academic literature on *program evaluation* were being applied; (2) consider whether the impact evaluation, together with information on costs was being used to properly assess cost-effectiveness; (3) provide a realistic assessment of the scope for incorporating existing methodological advances to help improve those evaluations; (4) provide an overall assessment of the evaluation, particularly with regard to its usefulness for policy makers in informing value for money assessments.

We agreed that the study would cover a selection of reports from the following four policy areas: active labour market interventions, business support policies, education policy and urban and regional economic policy including regeneration (henceforth referred to as 'spatial policy'). This entailed the use of reports from the Departments for Business, Innovation and Skills,

Communities and Local Government, Education and Work and Pensions (and their predecessors). Three central factors lay behind our decision to focus on these specific areas:

1. The policies in these areas are generally 'targeted' differently. Spatial policy targets areas, business support targets private sector firms, active labour market policies target individuals and education policy targets a mix of individuals and organisations. We thought this variation would help us illustrate and assess how government evaluation deals with a number of crucial methodological issues (data availability, establishment of counterfactual, appropriateness of outcome variables, etc.) discussed further below.
2. External perception of the quality of evaluation evidence varies markedly across these policy areas.
3. These are areas in which the project team have considerable experience (both in terms of academic and consultancy work).

Within these broad areas, we took several factors in to account when identifying particular evaluations to study: (1) the scale of the policy and of the evaluation itself; (2) the methodological approach taken, particularly with regard to the robustness of the approach; (3) the availability of comparable evaluations commissioned by overseas organisations.

The initial short-list was based on the research team's expert knowledge of policy in these areas plus searches of the relevant government department websites. The list covered programmes that ranged in scale and scope, and focussed on quantitative evaluations that could, in principle, lead to cost-effectiveness estimates. Note that we did not require the evaluations to have undertaken cost-effectiveness calculations. This was partly because it was felt that limiting the list in this way would prove overly restrictive and partly because suitable impact evaluations could have been used for internal cost-effectiveness assessments. Our preliminary list was refined after discussions with NAO (and further to email correspondence on the suitability of the selected evaluations between NAO and the relevant government departments). These discussions led to some minor modification of the initial list, although only to the extent that we felt these changes helped improve coverage in terms of the three criteria outlined above.

The final list, covering 35 evaluations, is provided in Appendix 1. Our assessment was based only on the published details of the evaluation reports rather than any additional information provided by departments. In many cases (particularly for active labour market programmes such

as the various New Deal programmes) the related evaluation literature is complex, multi-stage and multi-stranded, involving quantitative and qualitative analysis of the same programme over a number of years. In such cases we usually refer to final or synthesis reports in the list below while details of all documents consulted can be found at the end of the relevant study template. For reasons of objectivity, the list excludes evaluations carried out by members of the research team or their immediate colleagues and co-authors.

The rest of the report is structured as follows. Section 3 discusses and justifies the criteria used to assess the robustness of the evaluation reports. Section 4 details our findings of how the reports perform against these criteria, and makes a number of recommendations for improvements. Section 5 on 'when and what to evaluate' considers the question of the possible trade-off between robustness and the scope of evaluations. Section 6 concludes.

3. Criteria for evaluating the robustness of reports

Policy specific outputs (for example, the number of workers trained or firms assisted) are increasingly well monitored by governments. In contrast, many government sponsored evaluations that look at *outcomes* do not use credible strategies to assess the *causal* impact of policy interventions. The credibility with which these studies establish causality is (in our view) the crucial criteria on which evaluations aimed at providing estimates of impact and cost-effectiveness should be judged. By a causal estimate, the evaluation literature means an estimate of the difference that can be expected between the outcome for individuals 'treated' under a programme (i.e. affected by the policy), and the average outcome they would have experienced without it.[1] This question of attributing causality is of crucial importance in assessing the quality of evaluations and hence their usefulness as a basis for informing policy makers. Estimates of the benefits of a programme are of limited use unless those benefits can be attributed, with a reasonable degree of certainty, to the implementation of the programme in question. As emphasised by the literature on program evaluation and causal analysis more generally, solving this problem requires the construction of a valid counterfactual – i.e. what would have happened to the treated individuals had they not been treated under the program, an outcome that is fundamentally unobservable. Although these issues are traditionally considered for quantitative analysis, the same principles apply to qualitative analysis. It is

[1] For ease of exposition, we discuss the issues in terms of a policy that targets individuals, but the same issues arise when considering policies targeting areas, schools, firms or other organisations.

important to establish that individual opinions of benefits or perceptions of events genuinely relate to the policy implementation in order to draw any conclusions about its effectiveness, something that is nearly always overlooked in existing research.

The way in which this counterfactual is constructed is the key element of programme evaluation design. A standard part of this design is to create a comparator group of individuals not participating in or not eligible for the programme being evaluated. Outcomes can then be compared between the individuals subject to the intervention ('the treatment group') and similar individuals that are not exposed to the policy ('the control group').[2] Typically, outcomes are compared by looking at the differences in mean post-policy outcomes between these two groups. The assumption is that the post-policy outcomes in the control group provide an estimate of what would have happened to the treatment group in the absence of the policy. The challenge to effective programme evaluation is to ensure and demonstrate that this assumption is plausible, given theoretical reasoning, the institutional context and the evidence in the data.

Standard regression analysis can go some way in achieving this, by statistically 'controlling' for differences in characteristics between the treatment and control groups. However, this method can only control for 'observable' factors on which researchers have data and imposes some potentially quite restrictive ('functional form') assumptions about the way in which these observed characteristics affect the outcome in question. Modern programme evaluation techniques try to go further, by seeking to control for unobservable factors for which the researcher has no data or by relaxing the functional form assumptions.

A number of standard methods are used in the literature to overcome these problems, and they can be grouped roughly into four categories (although technically these are all overlapping):

- randomly assign units to the treatment and control groups as part of the programme design ('randomised control trial') so that on average the control and treatment group characteristics are the same;

[2] In situations where the intensity of treatment (e.g. the size of a grant) can vary it may be necessary to identify multiple groups receiving different treatment intensities. The difficulty then is in identifying groups that receive different treatment intensity but who would have experienced identical outcomes in the absence of treatment. In the main, for this report, we focus only on binary (policy on, policy off) treatment effects but many of the points we raise carry over to situations where the intensity of treatment can vary.

- use non-random treatment and control groups but subtract any pre-policy differences in outcomes in these groups from the post-policy differences ('difference-in-difference');
- focus only on differences in outcomes between units in the treatment and control groups that can be considered as randomly assigned, even if the groups as a whole are not. This can be done in two ways, by identifying a specific variable that predicts group assignment but not outcomes ('instrumental variables'), or by controlling for characteristics that predict group assignment and outcomes ('control function' approaches, including Heckman selection models, fixed effects models, and standard regression analysis);
- make the control group look more like the treatment group in terms of the observable characteristics of the members, by sampling a subset of the control group, or by weighting some members more than others, and then compare outcomes for this subset post-policy ('matching').

In the ideal setting, the programme would be designed with evaluation in mind and would contain an element of randomisation in the way individuals were made eligible for treatment. Failing this evaluations always face limitations on how certain they can be that any reported estimates are causal, because it can be difficult to construct a credible control group if evaluation has not been considered at the time the policy was devised. For example, if a policy is implemented at a national level, it will not always be possible to construct a valid counterfactual. Also, if policy targets all 'poorly performing' individuals, there may be no appropriate control group that is not subject to the policy. Another very difficult situation arises where individuals are targeted for a policy based on rather loose criteria. In this case, it is difficult to be confident that any control group is truly similar to the treatment group before the policy is introduced.

When selection is a problem and where randomised control trials are not an option, there are various statistical techniques that can be, and are used to, address this problem. Two examples are difference-in-differences and propensity score matching. Difference-in-difference techniques compare outcomes in an appropriately defined treatment and control group before and after the policy is implemented. The idea is that the change in the outcome for the control group between the pre and post policy periods estimates the change that would have occurred in the treatment group if the policy had not been introduced (i.e. the counterfactual). Therefore, subtracting this

change from the pre to post policy change in the outcome for the treatment group generates an estimate of the impact of the policy. The crucial assumption therefore is that the outcomes for units in the treatment and control groups would have followed the same trends over time in the absence of the policy. Although this assumption is ultimately untestable, some evidence can be gained by looking for differences in the pre-policy trends. It is not technically difficult or sophisticated to show pre-policy trends across treatment and control groups, if sufficient pre-policy data is collected, and there are variants in the difference-in-difference design that control for pre-policy differences in trends.

One of the most popular solutions to resolving treatment and control group differences (reflecting what was fashionable during the period when many of the evaluations we reviewed were undertaken) was matching. Matching involves pairing treatment units with control units that have similar, or identical, observable characteristics (implying that only treatment and control units which have some overlap in the available characteristics – i.e. 'common support' – are used). The approach is not unlike basic (OLS) regression analysis, except regression is typically more restrictive in the way it controls for these differences in observable characteristics between treatment and control units, because it assumes that the outcome is determined by a linear combination of the observable characteristics. Matching relaxes this assumption, and makes more explicit which group the estimates apply to: usually, each treatment unit is paired with a matching control unit, to provide estimates of the 'average effect of the treatment on the treated' (i.e. the effect for units which have the characteristics of the treatment group). Both standard (OLS) regression analysis and matching rely on the assumption that observable characteristics (those available in the data) are sufficient to account for all differences between treatment and control units that are relevant to the potential outcome of the policy. If this (untestable) 'selection on observables' assumption is correct, and there are no differences between treatment and control groups in the unobserved characteristics that cause differences in outcomes, then the difference in the outcome variable between the treatment and control units can be attributed to the effect of the policy in question. If the assumption is violated, then there is a standard omitted variables problem, and matching estimates, like standard OLS regression estimates, are biased.

Structure of templates for review of evaluations

The templates that we used to structure our retrospective reviews were designed to allow us to assess these key elements of the programme evaluation design. The first part of the template was designed to provide an overview of the policy and the evaluation. From each of the evaluations we collected information on the ***policy objectives***; the ***scope of the evaluation*** report; the ***overall methodology*** and, more specifically, the methods used for the ***impact evaluation***.[3]

The next part of the template considered ***policy detail***. Given the dependence of good evaluation design on the policy context, a basic starting point for credible evaluation is the availability of sufficient detail on the policy recorded at all stages of the policy making process. For example was selection of projects competitive? How were decisions made? What is the location and timing of intended and actual expenditure? Such information is crucial for understanding the way in which the policy works and thus for the establishment of a credible counterfactual.

The next set of issues concern data availability and 'construct validity' – that is the suitability of the data and variables chosen by the researchers to represent the policy itself and the outcomes the policy was expected to affect. Our reviews sought to cover a number of considerations. Was data available for the appropriate observational units over an appropriate time scale? For example, was data available before and after the policy was implemented in the relevant school or area. If so, was this data used appropriately in assessing the impact of the policy (e.g. was the outcome variable used appropriate to capture the impact of policy). If not, could this data have been collected? We asked a similar set of questions around cost data. Was there sufficient data available on costs to allow the evaluation to assess cost-effectiveness? This information was recorded in the templates under the areas of ***data***, ***costs*** and ***outcomes***.

Next, we turned to the ***methodological details***. Our main concern was with questions of '***internal validity***'. These assess the extent to which we can be confident that the results found identify some causal impact of policy on outcomes, at least for the observational units covered in the study. In this regard, the 'gold standard' in programme evaluation is large scale Randomised Control Trials outlined above. This ensures that a treatment and control group are similar at baseline. Of course this, by itself, is not sufficient for a top quality evaluation (other

[3] Note that the text highlights the headings that we use in our template in bold italics.

criteria would include suitably defined outcomes; appropriate duration; dealing with attrition; generalisability, lack of contamination effects). However, it ensures that the most basic requirement of evaluation is met: that there is an appropriate treatment and control group. There are situations when randomised control trials are either not possible or not appropriate. In this case, the methods listed above provide a starting point for other good evaluation designs that can be employed. However, it is not enough to simply use these methods without regard for the plausibility of assumptions in different contexts.

In our review, we consider whether, given the context and data availability, an appropriate methodology was implemented (e.g. difference in difference, propensity score matching). If so, was the methodology implemented appropriately? Were there any robustness tests done to test the adequacy of the methodology and the sensitivity to different estimation methods and assumptions about how to construct the counterfactual? For example, were 'placebo tests' used, based on other time periods or places where the policy was not in operation (the terms 'placebo' or 'falsification' test are used for demonstrations that the research design finds no policy impact in situations where no effect is expected e.g. when, or where there was no policy intervention)? Were balancing tests (including assessment of pre-trends) carried out to show that agents or areas in the treatment group exposed to a policy programme were comparable, in the pre-policy period, to those in the control group? Finally we turned to questions of '***inference***'. How precise are the estimates and how confident can we be that the results obtained have not occurred purely by chance as an artefact of the sampling process (i.e. have issues of statistical significance been properly addressed)?

Having assessed the internal validity of the evaluation we then turned to questions about '***external validity***'. This addresses concerns about whether the evaluation satisfactorily captured the impact of the policy for observational units not directly considered in the evaluation. This might involve the scaling up of results from samples to the national population (as happens a lot in labour market programme studies). There is also the question of how the estimated effect of the policy should be interpreted if the impact varies across different sub-groups of observational units. For example, even if the evaluation succeeds in providing causal estimates for the impact on units in the lowest income decile, what does this tell us about the impact if the policy is scaled up to cover other groups? The programme evaluation literature defines a range of alternative estimates which have different interpretations, and it is important that evaluations

are clear about what is being estimated e.g. the average effect of the treatment on the treated (the average causal effect on the group participating in the programme), the average treatment effect (the average causal effect on the population), or the intention to treat effect (the average effect on those offered an intervention, even if not all of these take up the offer or comply with it). Alternatively, questions of external validity could relate to assessing the wider and 'general equilibrium' impacts, for example whether the measured benefits of the programme were truly additional rather than displacement from other areas or agents, or whether the programme had spillover effects on those not included in the study. Lastly, external validity applies to questions about timing, for example whether the short run effects estimated by a study within a year of the policy would generalise to longer time horizons.

Any value for money assessment is only as good as the underlying impact evaluation. This is why we focus so much attention on the latter. But our review also considered the way in which ***cost-effectiveness*** was assessed. For example, were the estimated impact effects used to conduct a cost-effectiveness or benefit-cost analysis and if so how was this implemented? In cost-effectiveness analyses, was the effectiveness compared with other potential policy scenarios? Were appropriate shadow prices and discount rates applied in any cost benefit analyses?

Our reviews conclude with an ***overall assessment*** of the evaluation. We did not attempt to consistently address the details of the commissioning, length of time available for the study, etc., except to the extent that this was mentioned in any of the reports. Instead, we focused on the broader issue of the quality of the evaluation. Finally, we considered whether the resulting evidence was useful for policy makers in assessing the impact of the policy and whether or not it represented value for money. To help with comparability across policies and policy areas, we give each report a score based on our assessment of the quality of the impact assessment (see Box 1). The score is intended to be indicative of overall quality of the research design. However, this mainly relates to questions of internal validity. As a result it necessarily abstracts from a number of details, and we do not view it as a substitute for the more careful analysis presented in each of the templates.

Box 1: Overall scoring of evaluations

Our review adopts a scale based on the Scientific Maryland Scale (SMS) to provide a general indication of the reliability of the research design in providing estimates of the 'causal' effects of the policies being evaluated. This scale was developed by Sherman et al (1997) as part of a review of evaluations of policies targeted at crime reduction (summarised in Sherman et al 1998).[4] The SMS is a numerical scale ranging from 1, for studies based on simple cross sectional correlations to 5 for randomised control trials. The ranking implicitly indicates how effective the research design is in constructing a valid counterfactual for the policy intervention, and hence how reliably the estimated effects can be attributed to the policy in question. The scale relates mainly to questions of internal validity, rather than external validity and generalizability. The authors of the original scale did not provide a single precise definition of what type of study falls in each category, but provided a number of indicative descriptions. The following is our interpretation based on Sherman et al 1997, 1998).

Level 1: Correlation of outcomes with presence or intensity of treatment, cross-sectional comparisons of treated groups with untreated groups, or other cross-sectional methods in which there is no attempt to establish a counterfactual. No use of control variables in statistical analysis to adjust for differences between treated and untreated groups. We include qualitative studies that elicit ex-post views about a policy in this category, as well as the most basic quantitative evaluations.

Level 2: Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention ('before and after' study). No comparison group used to provide a counterfactual, or a comparator group is used but this is not chosen to be similar to the treatment group, nor demonstrated to be similar (e.g. national averages used as comparison for policy intervention in a specific area). No, or inappropriate, control variables used in statistical analysis to adjust for differences between treated and untreated groups.

[4] These two reports are available at https://www.ncjrs.gov/pdffiles1/Digitization/165366NCJRS.pdf and https://www.ncjrs.gov/pdffiles/171676.PDF.

Level 3: Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (e.g. difference in difference). Some justification given to choice of comparator group that is potentially similar to the treatment group. Evidence presented on comparability of treatment and control groups but these groups are poorly balanced on pre-treatment characteristics. Control variables may be used to adjust for difference between treated and untreated groups, but there are likely to be important uncontrolled differences remaining.

Level 4: Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (i.e. difference in difference). Careful and credible justification provided for choice of a comparator group that is closely matched to the treatment group. Treatment and control groups are balanced on pre-treatment characteristics and extensive evidence presented on this comparability, with only minor or irrelevant differences remaining. Control variables (e.g. OLS or matching) or other statistical techniques (e.g. IV) may be used to adjust for potential differences between treated and untreated groups. Problems of attrition from sample and implications discussed but not necessarily corrected.

Level 5: This category is reserved for research designs that involve randomisation into treatment and control groups. Randomised control trials provide the definitive example, although other 'natural experiment' research designs that exploit plausibly random variation in treatment may fall in this category. Extensive evidence provided on comparability of treatment and control groups, showing no significant differences in terms of levels or trends. Control variables may be used to adjust for treatment and control group differences, but this adjustment should not have a large impact on the main results. Attention paid to problems of selective attrition from randomly assigned groups, which is shown to be of negligible importance.

4. Our findings

From the start, NAO were clear that this was not intended to be a peer-review of the individual reports, their commissioners or the groups responsible for undertaking them. Many of these reports may have been peer reviewed by the departments concerned and have been commissioned in circumstances about which we know little. Instead, the focus is on drawing

lessons from the cross-section of studies. Individual analysts working within government departments would, we assume, have little time for such comparative work – especially outside their own areas. We were all agreed that the study needed to be conscious of the trade-offs faced by policy makers between the robustness of the evaluation findings, the cost of improving evaluation and the wider demands placed on evaluation during the policy making process. That said our reviews identified a number of areas where practice could be consistently improved at very low cost. We consider these in detail in this section and provide some initial ideas on how these issues might be addressed.

Scope of the evaluation and provision of technical details

Most reports that we considered have several strands including quantitative and qualitative components. They vary in the extent to which these strands are integrated into one report. Some of the reports only give the findings of the quantitative study, whereas qualitative findings are reported separately. In other cases, they are brought together in one long report. In some cases, such as when a programme (e.g. Business Link) is being re-evaluated following a change in delivery model this may make sense because the various components are complementary. In other situations this is more problematic. One particular issue that we would highlight is the extent to which detailed analysis of programme delivery is presented alongside the impact evaluation and cost-effectiveness components. This creates several problems. The target audience for these two reports are often different both in terms of interest and expertise. This often results in reports that appear to have been written for the 'lowest common denominator' – i.e. in such a way as to be comprehensible to someone who has no detailed knowledge of the policy and no training in programme evaluation. Such reports are often very long, but despite this provide insufficient detail for the specialist to assess the quality of the impact and cost-effectiveness components (e.g. volumes 1 and 2 of the national report on the impact of Regional Development Agencies run to 545 pages, but understanding the additionality assessment requires the reader to refer to other reports).

Some reports are accompanied by a published technical appendix that provides the necessary details. However, a surprising number of the reports we considered do not provide a technical appendix or, if they do, it is still written in a way that is aimed at a non-specialist audience. For example, in the areas of business support and spatial policy not one of the reports we looked at

provided a publically available technical appendix that was of sufficient quality to allow us to adequately assess the methods applied. In contrast, most of the education and labour market evaluations provided reasonably detailed technical appendices. We were also surprised by the extent to which practice on this varied within departments depending on the evaluation. For example, in the area of spatial policy, the technical appendix for the evaluation of the Neighbourhood Renewal Fund provided considerable technical detail for the modelling of neighbourhood transitions, while that of the Local Enterprise Growth Fund evaluation did not. Interestingly, the methodology adopted in the latter evaluation appears more appropriate than that used in the former, highlighting the fact that the lack of a technical appendix is not necessarily indicative of a weaker methodological approach.

A second issue to consider is the extent to which the same external organisations are capable of delivering *both* components – i.e. the detailed analysis of programme delivery and the impact and cost-effectiveness evaluation. Some of the larger evaluations deal with this problem by letting contracts to consortia (Education Maintenance Allowance; Every Child a Reader, Employment Retention and Advancement Demonstration and the Pathways to Work evaluation). In other situations, one research organisation has sub-contracted elements to other specialists (e.g. Centre for Research in Social Policy, who contracted the impact evaluation to a US research team for the New Deal for Disabled People evaluation). Sometimes, the same organisation handles both components (e.g. Small Firms Loan Guarantee). This is clearly desirable in situations where the analysis of programme delivery complements the impact evaluation. In other situations, however, it is unclear that the two components should be contracted at the same time, or even to the same organisation. For example, for many policies, impacts may take some time to be felt, but questions around delivery need addressing in the short term. In such situations, impact and cost-effectiveness evaluations may not be feasible on the same time scale (the Mixed Communities Evaluation provides a good example). The question of contractor capacity is a difficult one for us to address here. But we are concerned that some of the reports we reviewed suggest that the requirements on the assessment of programme delivery may rule out contractors who would be better placed to undertake the impact evaluation components of the assessment.

These problems are exacerbated because reports that are *less* careful about identifying the causal impacts of policy are often willing to make much broader claims about the impact of a

policy (and how that impact was achieved). As a result, policy makers face a difficult trade-off when trying to decide how to evaluate policies. Wide ranging 'evaluations' that are less careful about causality appear on the surface to provide more information as an input in to the policy making process. Taken at face value, such evaluations allow policy makers to both assess value for money and make changes to policy, while appearing to take in to account evidence about the impact of the policy. However, without paying careful attention to causality, the information these evaluations provide is of dubious value. In contrast, empirical research in the programme treatment effects tradition is often circumspect and makes fairly narrow claims about whether the policy has a causal impact (and then, sometimes, only for a particular part of the population depending on the methods used).

Recommendation 1: A technical appendix, written for a specialist audience, should be a core component of every impact evaluation. This technical appendix should provide sufficient detail on all aspects of the impact evaluation to allow external assessment by a specialist in the field. It should be easily accessible, usually in electronic format, and available on the departmental web page alongside the overall report.

Recommendation 2: The full scope of the evaluation should be clearly identified at the start with more care taken to distinguish between the analysis of programme delivery (process evaluation) and the assessment of impact and value for money (impact evaluation). Based on this, the range of expertise needed to carry out all elements of the evaluation should be identified, so that the approach to the selection of contractors (or internal evaluators) ensures that they have capacity to carry out all elements of the work to appropriate standards.

Appropriate data and outcomes (construct validity)

The reports that we considered covered a wide range of policies and policy areas. In some instances, the policy objectives were clear enough that it was relatively easy to identify outcomes that could be considered to assess the impact of the policy (Education Maintenance Allowance; Every Child a Reader; Every Child Counts; Activity Agreement Pilots, New Deal programmes, Pathways to Work). In other areas, however, the eventual policy objective was more difficult to translate. Indeed, our review of evaluations highlighted a surprising number of

reports where information on the details of the policy was not available (at least to the organisation undertaking the evaluation). For example, in the case of the evaluation of the National Strategy for Neighbourhood Renewal, the team evaluating the education component did not know how the education funding had been used even at a basic level (e.g. targeted at schools or individuals; pre-school, primary or secondary). This made it difficult to know which outcomes were really appropriate for evaluating the policy. Another example is provided by the Local Enterprise Growth Initiative which aimed to 'release the economic and productivity potential of the most deprived local areas across the country'. In these cases, where policy objectives are not clearly specified, or are not specified in terms of outcomes that are easily measured, evaluation may need to assess the impact on intermediate outcomes (in the case of Local Enterprise Growth Initiative, e.g., on enterprise and investment). When data on these intermediate outcomes is not available reports often need to use proxy variables (in the case of LEGI, e.g., the number of new businesses in an area may act as a proxy for 'the release of economic potential'). This raises a specific difficulty – if the evaluation shows no impact on the proxy it is easy for the report to conclude that this is, anyhow, an imperfect measure of policy impact. In contrast, positive effects are usually taken at face value as identifying the impact of the policy in question. Evaluation of the impact of spatial policies on educational outcomes (New Deal for Communities and Neighbourhood Renewal Fund) provides an example of a report where this problem arises – positive coefficients on a limited number of proxies are highlighted over zero coefficients on many more proxies (some of which are arguably more relevant to the programme).

One problem with the use of a control group can be in getting data for organisations or individuals who do not participate in the programme. Bespoke surveys of both participants and non-participants can substantially add to costs, particularly when an extensive survey is already undertaken to deal with the programme delivery components of the report. In the labour market evaluations these restrictions have to some extent been relaxed through the use of administrative data, although this is often combined with additional bespoke surveys, because administrative data does not provide a rich set of controls (New Deal for Lone Parents, New Deal for Disabled People) – and similarly in education, through the National Pupil Database. There appears to have been far less use of such data in the area of spatial policy and business support. This is despite the fact that such data sources are increasingly available (to academic researchers through the Secure Data Service at Essex, and for non-academic researchers

through the ONS Virtual Micro Laboratory). In some circumstances, data available from existing sources may only proxy for intended outcomes (as just discussed, e.g. in the case of Local Enterprise Growth Initiative where the number of businesses may be used as a proxy for the intended outcome of 'releasing economic potential). In these cases, there is still a case for value for money calculations to be based on these proxy variables used with a case-control group in a robust analysis. In most cases, the alternative strategy is to cover a wide range of outcomes based on surveys of participants and self-assessed additionality. Our review suggests that these approaches are far more prevalent in the areas of business support and spatial policy – see for example, reviews of Regional Selective Assistance, Small Firms Merit Award for Research and Technology/Support for Products Under Research, Manufacturing Advisory Service and Business Links. This raises the possibility that departmental 'norms' may play a role in influencing the extent to which such approaches are used (rather than a detailed consideration of what might be most appropriate for the policy at hand).

Recommendation 3: If the study needs to use a proxy outcome variable as a result of data availability, then negative or zero outcomes for that proxy outcome should be treated symmetrically to positive outcomes.

Recommendation 4: Departments should review the extent to which existing administrative data, possibly on proxy variables, could be used more effectively to evaluate the impact of policy in a case-control setting.

Use of a counterfactual for impact evaluation

As discussed above, in our view, any impact evaluation (and subsequent value for money calculation) requires construction of a counterfactual. A majority of the reports that we have considered, 27 in total, use a treatment and control comparison as a central component of the evaluation. Once again, however, there was considerable variation both between and within policy areas, with education and active labour market policy evaluations far more consistent in the use of a (valid) counterfactual than spatial or business support evaluations. It is important to remember, however, that our sample is skewed by the fact that we have considered reports which provide, or present findings that could be used to provide, a value for money assessment. That said a significant minority of the reports that we reviewed, 6 in total, do not use a control

comparison group or any other method for establishing a counterfactual. Some of these reports still refer to their findings as providing an impact assessment and report value for money estimates. See, for example, the evaluation of Regional Selective Assistance (1991 to 1995).

We recognise that it may not always be possible to construct a valid control group, or find any other way of establishing a credible counterfactual. The difficulty in devising appropriate estimation strategies can depend on both data availability and the nature of the government programme. This may partly explain why insights from the programme evaluation literature take centre stage in many recent evaluations of labour market programmes carried out for and by the Department of Work and Pensions (although not in all cases e.g. the Fair Cities Pilot), but they take a less prominent role in evaluations of business support or spatial policy. In short, the scope for informative impact evaluation is not necessarily uniform across policy areas. That said, differences in the inherent difficulty of assessing different types of policy interventions cannot explain all of the variation in the quality of evaluations that we identify below.

With this caveat in mind, in the following sections we focus on reports which make use of a counterfactual group for impact analysis and turn to questions of the quality of this analysis. We return to consider the issue of the lack of a suitable counterfactual in section 5 which discusses the question of '*When and what to evaluate'*.

Recommendation 5: More care should be taken to distinguish carefully between impact assessments that are based on the use of a counterfactual versus other approaches (for example based on self-assessment of additionality). It should be recognised that the use of a control group, or other counterfactual is a necessary, although not sufficient, requirement for robust impact assessment (and value for money calculations based on that impact assessment).

Internal validity, inference and external validity

As discussed above, by internal validity we mean the reliability of the evaluation in providing causal estimates of the impact of the policy on the group being studied. There are four over-arching questions: 1. Was the research design appropriate? 2. Was the design implemented properly? 3. Were the assumptions behind the research design tested and demonstrated to

hold? 4. Did the report explain the methods clearly enough to allow assessment, and were any limitations and caveats highlighted.

The reliability of the research design is limited by the design of the policy being evaluated. The design that offers the most straightforward (but potentially costly) route to a reliable evaluation is a randomised control trial, but clearly this is not possible unless built into the programme delivery at the outset. In the US, is it fairly common practice to commission randomised control trials (e.g. Head Start shares similarities with Sure Start in the UK). In the UK, randomised control trials are beginning to be used more in education, albeit using relatively small sample sizes (e.g. the Teens and Toddlers Programme; Every Child Counts), and in active labour market policy (e.g. Employment Retention and Advancement Demonstration). However, in the UK, trials in these policy areas are currently the exception not the rule.

Even in this best case scenario, it is rarely possible to design or implement randomised control trials perfectly. One problem is the temptation to reduce costs of randomised control trials by making them too small and too short-term. For example, some randomised control trials use the minimum sample sizes required (theoretically) to detect quite large impacts at low levels of statistical significance. The scale of RCTs need to be large enough to detect small-medium sized effects unless there is very good reason to expect large affects a-priori. For example, the Teens and Toddlers at Risk sample was chosen based on the minimum size necessary to detect a halving of the proportion not using contraception, at a 5% statistical significance level. Conducting randomised control trials at small scales risks wasting the resources devoted to evaluation by producing estimates of effects that are too imprecise to be useful, either because they report significant positive impacts that have occurred purely by chance, or because they miss impacts which although small in magnitude would imply cost effective policy when rolled out to the population. Even if sample size is sufficient, there are likely to be issues of drop out and non-compliance by participants, and methods need to be adopted to compensate for these problems. Furthermore, it is very important to take account of potential spillover effects in how the RCT is designed. The few RCTs we evaluated were good at addressing some of these problems. For example, there was a separate report produced to assess potential design problems that were discovered in the Employment Retention and Advancement Demonstration; The Teens and Toddlers programme evaluation included a reserve pool to be used to replenish the treatment group if participants dropped out. However, the design of the Teens and Toddlers

programme deliberately overlooked potential spillover effects because of the costs of doing a study that would have accounted for this (using a clustering design). With regard to the RCT done as part of the 'Every Child Counts' evaluation, it was unfortunate that the design did not allow for evaluation beyond 12 weeks (which was instead assessed using a much weaker methodology).

In the absence of explicit randomisation, studies are forced into adopting other methods to create comparison groups and counterfactual outcomes. Researchers and policy analysts will have different views on the likely effectiveness of these methods in general, so it is imperative that evaluations carefully describe their techniques and demonstrate in the reports that the method is effective on a case-by-case basis.

In the absence of a randomised control trial, the programmes that were rolled out sequentially across different geographical areas or piloted in a subset of areas provide good potential research designs, as there are clearly defined groups that are eligible and not eligible for programme participation according to the programme design (rather than on the basis of individual decisions to participate in the schemes). Many studies in the education and labour market areas made good use of this kind of design (e.g. Educational Maintenance Allowance , the Pathways to Work and Job Centre Plus evaluations). In contrast, we found no examples in the areas of business support or spatial policy that explicitly made use of staggered policy implementation, even though a number of policies in this area had similar characteristics (for example, the Single Regeneration Budget programme involved six rounds of spending over a number of years suggesting that projects funded in later rounds could have been used to construct a suitable control group for projects funded in earlier rounds).

In cases where a programme was national, and there was no piloting or staggered rollout, many evaluations fall back on comparing individuals or firms that selected themselves into the programme voluntarily or were selected in by external agencies, with others who did not participate in the programme. For example, in several of the education reports the control group is made up of people or schools who chose not to participate in the treatment that is being evaluated (e.g. National Citizen's Service Pilots; Social and Emotional Impact of Learning programme) or not chosen by their Local Authority (Key Stage 2 Careers-Related Learning Pathfinder Evaluation). This problem is not restricted to education. For example, in the

evaluation of Regional Selective Assistance and Selective Finance for Investment in England, the counterfactual is taken from a group of firms who did not receive assistance, while the analysis of Small Firms Merit Award for Research and Technology/Support for Products Under Research uses unsuccessful applicants. In a number of active labour market policy evaluations (e.g. New Deal for Lone Parents, Work Based Learning for Adults, New Deal for Disabled People) individuals who volunteered for participation in training or job search support schemes were compared with individuals who did not. In the context of treatment/control group based policy evaluation, this type of research design would be considered the option of last resort. In such cases, there are implicitly important differences between the treatment and control groups in terms of motivation and their pre-policy history. Indeed, many studies documented these differences (see, e.g., evaluation of Regional Selective Assistance and Selective Finance for Investment in England). These differences between treatment and control groups raise questions about the effect that selection is having on the results.

In such circumstances, the design of the policy means that it is very difficult to construct an appropriate counterfactual which allows the selection effect to be fully addressed. In these cases, the limitations need to be acknowledged and the resulting bias carefully discussed. Generally speaking, it may be preferable to set up the analysis so as to ensure that selection bias is hopefully working in only one direction. For example, if treatment is only based on the quality of the application for funding (e.g. as may be the case for Local Enterprise Growth Initiative) then comparing successful to unsuccessful applicants might arguably produce an upward bias. If the direction of bias on the policy coefficient is upward, then zero coefficients are still informative about the lack of impact of the policy. Similarly, in situations where selection is likely to generate downward bias (e.g. when treatment depends on some assessment of need), positive coefficients are likely to underestimate the impact of policy.

In contrast, when multiple biases work in opposite directions estimated coefficients are almost impossible to interpret (unless there is strong reason to think that a particular selection problem dominates). The evaluation of Regional Selective Assistance provides a good example when funding is given to firms that successfully apply for funding (possibly generating an upward bias) to safeguard jobs (possibly generating a downward bias if firms are struggling). Other Regional Selective Assistance grants are given to firms who can demonstrate that the project will create jobs (which possibly creates an upward bias). In the absence of further information, these

different biases are impossible to disentangle a-priori. It is, however, often possible to explore these issues by observing how sensitive estimates are to changes in the way the treatment and control groups are specified and made comparable, or to the introduction of additional control variables in a regression context. Very few of the studies we reviewed do this (exceptions include a supplementary report for the Employment Retention and Advancement demonstration that provides a range of sensitivity tests, and the New Deal for Lone Parents re-evaluation which explores the sensitivity of the results to a number of different specification changes).

Of course, it is much better to set up an evaluation that properly deals with selection rather than have to interpret findings which could well be biased. Here, our selective review of evaluations suggests that there is a danger of setting the bar too low and failing to keep pace with international standards in programme evaluation. For example, a recent World Bank report reviews international impact assessments of youth voluntary service programs, but the interim report of the National Citizen Services Pilot does not suggest that issues around best practice (for example, as discussed in this World Bank report) have been incorporated into their design. Our review suggests that this is a problem that holds more widely.

As discussed in section 3, when selection is a problem and where randomised control trials are not an option, there are various statistical techniques that can be used to address this problem. One possibility is to use difference-in-difference based on *changes* in outcomes for a treatment and a control group. Because the validity of this technique rests on the (untestable) assumption that the treatment and control group would have followed the same trends in the absence of policy it is good practice to test whether this assumption at least holds pre-policy. Indeed such tests were often implemented in the labour market programme studies we reviewed (see, for example, the Job Centre Plus evaluation). Yet, this basic check appears to be overlooked in a number of the evaluations we have considered, particularly in the areas of spatial policy and business support.

A second reason to expect violation of the difference-in-difference design assumptions is that other interventions were implemented differentially across the treatment and control units considered over the same time period as the intervention itself. This is potentially a general problem across many of the evaluations, because almost all of the evaluations were taking place over the first decade of the 21st century in the same geographical area (Britain) and yet

almost all studies ignored the possibility of contamination from other programmes. Labour market evaluations (e.g. New Deal) acknowledged that they were carried out in an environment with many other simultaneous interventions (often on the same groups of people!). Some made efforts to exclude individuals participating in other programmes (e.g. Work Based Learning for Adults) although this in itself can lead to the control group being a selected sample, with consequences for internal validity. In general, however, it was often not completely clear in the evaluations how to interpret the estimated impact of policies that were being introduced simultaneously with other policies.

Another solution widely used to resolve treatment and control group differences was 'matching' or OLS regressions. Examples in the areas of education and active labour market policy include New Deal for Lone Parents and New Deal for Disabled People and the Education Maintenance Allowance evaluation. Once again, however, these methods are far less common in spatial policy or business support with matching used in two reports for the former and no reports for the latter. As discussed in section 3, the validity of these techniques rests on the (untestable) assumption that observable characteristics (those available in the data) are sufficient to account for all differences between treatment and control groups that are relevant to the potential outcome of the policy.

In some cases, for example where a programme roll out generates groups of eligible (treatment) and ineligible (control) groups, matching or OLS regression methods are potentially appropriate, because there are likely to be treatment units (individuals or firms) that are very similar to the control units. The latter are simply excluded due to programme availability, not through personal choice. In other cases, where programme participation is voluntary and likely dependent on unobserved personal characteristics, or based on selection of participants by other agencies, matching on observable characteristics is unlikely to provide a very satisfactory solution. Nevertheless, this method was adopted in many of the studies we reviewed (e.g. New Deal for Disabled People, New Deal for Lone Parents; National Citizen's Service Pilot).

In these circumstances, best practice suggests reports should present results both with and without the correction for selection so that the extent of the selection problem can be considered. If coefficients are very similar but it is believed that selection affects are likely to be strong, this also raises questions about the validity of the approach used to address selection.

This would usually mean presenting simple OLS estimates for comparison to the results from more sophisticated techniques. Again, we were surprised to the extent to which this did not happen in the evaluations that we have considered.

We also found instances in which propensity score matching was simply not used correctly. For example, the educational component of the National Strategy for Neighbourhood Renewal used the approach to trim the sample of control schools, but not to compare treatment and control schools that exhibit 'common support' (that is to say, ignores schools that were very likely or unlikely to be treated so as to focus comparison on treated schools which conceivably could have been untreated and vice-versa).

A number of the reports we reviewed use techniques that have been widely superseded in the evaluation literature. Interestingly the 'misuse' of these techniques often occurs repeatedly in different evaluations undertaken by a particular contractor or in evaluations for a particular department. For example, evaluations of business support schemes Regional Selective Assistance and Small Firms Merit Award for Research and Technology/Support for Products Under Research (SMART/SPUR) see contractors use the Heckman selection correction for different policies. Although this is a useful method, it has been recognised for some time (since at least the late 1990s) that the results it produces are sensitive to the specification of which characteristics determine treatment group assignment, and a careful case must be made for the exclusion of one or more of these characteristics from the set of characteristics which are allowed to affect the outcome. Unfortunately, none of the evaluations that we considered that applied this technique discussed the handling of these issues clearly, if at all (e.g. Regional Selective Assistance, SMART/SPUR). Such problems can also be perpetuated across sequential analysis of long lasting policies if contractors are restricted to approaches that make their results comparable to earlier reports. This happens, for example, with the review of Regional Selective Assistance where the same process (using no-control) group is deliberately replicated in later reports (for more details on both these concerns see the relevant templates in the appendix).

The next set of questions concern inference, i.e. the statistical reliability of the coefficient point estimates. Many reports, particularly in the areas of business support (4 in total) and spatial policy (6 in total), provided no indication of the statistical significance of the estimates. Even

those that did (e.g. nearly all of the labour market programme evaluations reviewed) were not always clear about the way these were estimated, and the assumptions behind these estimates. For example, when programmes are delivered at an area level, or by jobcentre plus offices, there are potential correlations between unobserved characteristics across neighbouring units or units being treated by the same offices, which could lead to biases in standard errors, confidence intervals and tests of statistical significance. Statistical significance tests of difference-in-difference estimates and other methods that follow observational units over time are also prone to problems caused by correlation in these unobservable factors over time. There are methods for estimating standard errors to allow for these types of problem (clustering) which are routinely applied in academic work, but these were not applied in the evaluations we looked at (or may have been applied, but details were not provided).

Selection problems also complicate the interpretation of the treatment effects identified in the evaluation. In some cases the 'Intention to Treat' parameter is estimated (e.g. Education Maintenance Allowance; Teens and Toddlers Programme; Sure Start), which allows for the fact that some eligible people will choose not to participate in the programme or drop-out. This evaluates the effect of the programme on all eligible participants. The 'treatment on the treated' effect is harder to estimate because of self-selection into (or out of) the treatment among the eligible population. This problem is either not recognised in some studies or inappropriately overlooked (e.g. Activity Agreement Pilots). It is surprising how little information is given on exactly what impact is being estimated (intention to treat versus effect of treatment on the treated). There are some examples where both could have been reported – but the evaluation argues for, and only reports, one or the other (e.g. Activity Agreement Pilots).

Moving from the coefficient estimates to the total impact of the programme requires a number of decisions to be made about how to scale up numbers. Again, there seems to be considerable variation across departments in the way that this is done that are not purely attributable to the nature of the programme under study. For example Active Labour Market evaluations (e.g. New Deal for Disables People, New Deal for Lone Parents), are good at correcting for non-response and weighting up to national numbers but don't consider general equilibrium effects. These matters are occasionally considered in education evaluations (e.g. Education Maintenance Allowance). In contrast, evaluations of business support and spatial policy are less careful about how to aggregate up, but more conscious of general equilibrium effects, not that these are ever

very well estimated (see, e.g., Regional Selective Assistance evaluation discussions of displacement). That said, even in very good evaluations, one does not often get a picture of how the schools/individuals/areas that have been chosen as the target of various programmes compare to overall population, making assessment of external validity very difficult.

Finally, it should be emphasised that one of the major difficulties we faced in assessing the quality of the impact assessments concerned the details provided in reports on methods of estimation. A small number of studies use strict protocols in how the evaluation is reported (e.g. the randomised evaluation in Every Child Counts uses guidelines from the Consolidated Standards on the Reporting of Randomised Trials - http://www.consort-statement.org/index.aspx?o=2965). In many cases, however, too little information was provided to allow expert assessment.

Recommendation 6: Reports need to pay far more attention to the problems of selection and the extent to which this affects interpretation of the policy impacts.

Recommendation 7: Where statistical techniques are used to correct for selection, the report should provide both the corrected and uncorrected estimates to allow the extent of selection bias to be assessed. Results should also show the sensitivity to the inclusion and exclusion of different sets of matching or control variables.

Recommendation 8: The techniques applied should be appropriate for the policy issue in hand, given current knowledge in the programme evaluation literature. Furthermore, mechanisms need to be put in place to ensure basic mistakes are avoided in how techniques are applied. These might include mechanisms to ensure adequate training and up-dating of analytical skills for staff and appropriate internal or external peer review.

Recommendation 9: Estimates should be reported with indications of statistical significance, standard errors or confidence intervals, and the methods and assumptions used to estimate them.

Recommendation 10: The Intention to Treat parameter should be estimated as a priority because of fewer problems of selection bias. However, where it possible to estimate additional parameters (e.g. the Impact of Treatment on the Treated) in a credible way, then all estimates provided should be clearly defined.

Recommendation 11: Studies need to consider issues relating to external validity. At the very least, they should place their evaluations in a broader context by showing how the characteristics of treatment and control samples compare to the wider population of interest.

Recommendation 12: The technical report or appendix must give sufficient detail to allow a specialist to assess the approach taken in terms of internal validity, inference and external validity.

Cost-effectiveness

It is important to recognise that the extent to which cost-effectiveness is actually covered by the reports varies considerably and that this variation is likely to have been driven by the project specification. Many of the labour market evaluations contained cost-effectiveness or cost benefit calculations (in one case, Pathways to work, there was a separate report on the cost benefit analysis). In contrast, hardly any of the education projects contained cost-effectiveness calculations and we were informed by DfE that these calculations tended to be done in house based on the evaluations. There are, of course, good reasons why this may be the case. That said, it would clearly be desirable for the resulting cost-effectiveness calculations to be made available alongside the impact evaluation.

A major barrier to cost-effectiveness evaluation is the lack of systematic data collection on costs. This appears to be a specific problem for evaluations of spatial policies – particularly when delivery is 'devolved' to local government. See, for example, reports on Local Enterprise Growth Initiative and Neighbourhood Renewal Fund, for in-depth discussion of the problems faced in getting usable cost data.

In many of the cases where cost-effectiveness calculations are undertaken the approach taken is usually reasonably narrow. Adjustments on the cost side usually consider direct costs (adjusting for taxation and the costs of delivery) but ignore the costs to participants that would be needed for a full cost benefit analysis. On the benefits side, the expected duration of benefits receive some attention (e.g., in particular, in the business support studies on the impact of Regional Selective Assistance) but there is little consideration of more detailed timing and discounting (there were exceptions in the labour market programme evaluations, e.g. Pathways to Work, New Deal for Disabled People, although the New Deal for Disabled People long run benefits were estimated by predicting out of sample using methods that lacked credibility). Where general equilibrium effects are expected (e.g. for area based policies such as Regional Selective Assistance) further adjustments are often made for displacement and multipliers – although this is nearly always based on self-reported 'guesstimates'.

Many of the reports we have reviewed could conceivably place more emphasis on cost-effectiveness although we question the value of this until estimates of impact are improved. Indeed, in some circumstances we worry that the need to provide value for money estimates may distort the evaluation process. For example, in the SMART/SPUR evaluation (SMART is an acronym for Small Firms Merit Award for Research and Technology SPUR for Support for Products Under Research) the econometric impact evaluation suggests that the policy has no impact. But the value for money calculations use self-reported additionality which then give reasonable value for money figures. The reasons given for this relate to the statistical validity of the econometric results plus the fact that the evaluation was done 'too early' to capture the full effects (even though it is assumed that firms are accurately able to predict what these will be – which stretches credibility).

Recommendation 13: Where an impact evaluation is used as the basis for an internal cost-effectiveness assessment a report providing details of that assessment and its conclusions should be made available. It should be easily accessible, usually in electronic format, and available on the departmental web page alongside the overall report.

Recommendation 14: Cost-effectiveness calculations require data to be available on the pattern of spend (across individuals, activities, locations etc). In situations where such

cost data is not being systematically collected, this problem needs to be addressed urgently if policy is to be effectively evaluated in the future.

Recommendation 15: Unless the impact evaluation meets minimum standards, there is little point in doing a value for money calculation using estimated impacts from that evaluation. Such analysis is misleading. We recommend that value for money calculations that rely on estimated impacts are only conducted after a sufficiently robust impact evaluation that provides estimates that are credible and based on meaningful outcome measures. In some circumstances, the comparison of gross outcome to costs (assuming 100% additionality) may be useful to identify programmes that are particularly poor value for money.

5. When and what to evaluate?

The discussion so far raises the question of what policy makers should do in situations where a suitable control group cannot be identified. Our review of evaluations suggests that the usual solution in this situation is to use self-reported assessments of benefits (see, for example, evaluations of Regional Selective Assistance, or SMART/SPUR). There are several problems with this approach. First, this may give widely distorted assessments of impacts and, as a result, of value for money. At the very least, it is difficult to believe that such self-reported estimates form a valid basis for comparisons across different policy areas that use different types of intervention (e.g. of the value for money of labour market versus business support – as was done in the national evaluation of the Regional Development Agencies). Even within policy areas we know of no systematic analysis of how such self-reported assessments vary conditional on the characteristics of those being asked to provide the assessment. Say, for example, for some policy small firms report more additionality than large firms. We have no guide on the extent to which this reflects systematic differences in the tendency of small and large firms to report different additionality independent of what happens in reality. Similarly, do specific project managers (responsible for one component of delivery) tend to report higher or lower additionality than programme managers (responsible for multiple components)? What about civil servants working in central government as opposed to local government? Given all this uncertainty, it is hard to know what to make of self-reported assessments even for similar types of policies, unless we know that studies have broadly similar groups being asked about

the impact of the policy. This problem may be more acute in some areas (e.g. education, spatial policy and business support) than it is for individuals although, we stress again, that we have no evidence on which to assess this assertion. Similarly, we do not know how self-reported additionality varies with elements of project delivery (e.g. the quality of promotional materials) that may make no difference to additionality in practice. These are serious concerns and they are not adequately addressed in any of the reports that we have reviewed that use this approach.

One way to address this problem (of the lack of suitable control groups) is to move away from blanket evaluation of entire programmes and instead to focus only on those areas of the policy that are amenable to more rigorous evaluation. A number of questions may help identify situations in which such a strategy would be a better option:

1. Is impact evaluation appropriate? Several dimensions - magnitude of spend, likely cost of evaluation - are already considered by departments when deciding whether to undertake an evaluation. It should also be possible to identify situations where careful monitoring and analysis of process delivery may be more appropriate than impact evaluation. For example, the rationale behind the “Key Stage 2 career-related learning Pathfinder” was to improve the quality of careers-related information provided to pupils in primary school. Local Authorities (and their selected schools) were allowed to develop their own approaches. In this situation, monitoring and analysis of process delivery is very important for assessing the Pathfinder. However, impact evaluation is of questionable value (at least at an early stage) because (a) what schools are doing is not clear; (b) the counterfactual is not clear; (c) it is difficult to specify appropriate short-term outcomes that are quantifiable.

It is also difficult to conduct impact evaluation in the context of national programmes where take-up is very high because there are good reasons to think that the minority group of non-participants are somehow different from the majority group that participate. For example, in the Social and Emotional Aspects of Learning evaluation schools that choose not to adopt the programme will be different from schools that choose to take part in ways that are very difficult to capture. The same may apply in instances where eligibility is unrestricted and yet take-up of the programme is very limited.

2. Could focussing on a narrower set of outcomes allow more robust impact evaluation, possibly using administrative data? For example, given that the Inter Departmental Business Register provides good administrative data for covering 99% of UK economic activity would it be better for evaluations to focus on the employment impact of business support schemes? The better quality labour market evaluations have made extensive use of administrative data (especially linked DWP benefits and HMRC earnings data). The impacts of such a shift in focus may be substantial. For example one of the findings of the re-evaluation of the New Deal for Lone Parents evaluation by Dolton et al (2006) was that switching the evaluation to administrative data (plus some other refinements) resulted in estimates that were half those in the original evaluation carried out on survey data. In general, administrative data offers greater potential in terms of external validity (and sample sizes), the trade-off currently being less rich information about the characteristics of the units of analysis, since administrative data usually collects a limited set of such information. There is an argument here for collecting and making more information available from administrative sources for general evaluation purposes.

3. Can details of the policy be used to identify the impact on particular groups even if not on the treated population as a whole? One possibility is the use of eligibility rules to implement regression discontinuity designs. For example, when business support policies are targeted at small to medium size enterprises the restriction is usually implemented in terms of firm size (say, smaller than 250 employees). In these cases firms just above the threshold may act as a suitable control group for firms just below the threshold. Under certain conditions this approach gives a good estimate of the causal impact of the policy on firms close to the threshold. Depending on the set up of the policy (e.g. the extent to which firms manipulate the cut-off variable to become eligible for the policy) this may allow estimate of the impact of the policy even when firms select in to treatment (at least, that is, the impact on the treated). These effects can be compared to impacts for the treated group as a whole to give some idea of the extent to which selection biases those estimates. The threshold estimates may also be interesting in and of themselves in situations where changes to eligibility criteria are being considered. Similar strategies can be developed using, for example, test scores or geographical rules for eligibility (by looking at outcomes for those just inside the eligible area to those just outside). Academic research involving one of this report's authors is currently using such approaches to revisit the evaluation of the Local Enterprise Growth Initiative (which exhibits both a spatial and index of multiple deprivation 'discontinuity' created by eligibility rules).

The academic literature also makes increasing use of instrumental variable strategies to help solve the selection problem. For example, when changes to eligibility rules make some previously eligible firms ineligible, or vice-versa, those who have experienced a change to their eligibility can act as a control group for those who continue to be treated by the policy. Such an approach has recently been used by one of this report's authors to assess the impact of Regional Selective Assistance (by using changes in the UK map of eligibility). The academic literature has also made considerable progress in interpreting these estimates in situations where effects are not uniform across the treated groups.

If a programme is implemented nationally, or has strong self-selection into treatment it may be very hard to think of ways to undertake an effective evaluation of the overall effect of the policy. It is arguably a waste of time and resources trying to do impact evaluation for the overall impact of such policies in these circumstances, as findings will not be credible. Among the issues addressed so far are: 1) Self-reported assessments of additionality – we think these are highly problematic and relied upon too much in the existing reports; 2) Focussing on process, delivery and monitoring rather than impact evaluation; 3) Focussing only on particular outcomes for which we have good data; 4) Focussing on particular groups for which policy details allow effective evaluation. If the scheme is large, involves considerable expenditure and we want to know the *overall* impact, none of these solutions may be particularly attractive. Unfortunately, there is no 'magic bullet' solution, but making progress requires much more recognition of the fact that impact evaluation needs to be embedded at the start. If interest is in the overall impact of the policy on a range of outcomes then this means piloting the study. If interest is in identifying the impacts on some groups rather than others, then it may be possible to reflect this in policy design allowing identification of the effects using features of the programme evaluation literature. If policy is to do this it further needs to recognise importance of (1) not conflating with other policy interventions; (2) allowing time for effects to happen. This sounds difficult, but this goal has, at times, been achieved – even in the context of ambitious, national policies. For example, the Education Maintenance Allowance is an example of an education policy that allowed for careful evaluation in a treatment-control context over 3 years, even when it was being rolled out nationally (as the control areas were the last to receive treatment).

Previous experience, plus our review of the projects leads us to conclude that most official evaluations are far too short-term and that there is no mechanism for evaluating longer term impact (even if this is not funded as part of the project evaluation). In the context of a desire to thoroughly evaluate the cost-effectiveness of programmes, this is very short sighted. It may well be a consequence of various pressures within departments to come up with quick answers on impact. This is one of the reasons why critical appraisal of evaluations may need to come from outside particular government departments both before and after the commissioning process (as is common with major projects commissioned by the ESRC). Also, there is a strong argument to be made for funding a more select number of projects – to ensure high quality evaluation – rather than trying to evaluate a large number of projects. Currently, there is huge variability in the quality of projects (even within government departments). Even within quantitative studies, a narrower focus on very specific questions with a good methodology would be preferable to applying lots of different strands, where only some of this analysis is capable of giving useful insights ('Every Child Counts' is an example of a quantitative evaluation with a top quality component and other less useful components).

Recommendation 16: When robust impact evaluation is not possible it is important to recognise that commissioning an evaluation may not represent good value for money. In these circumstances process evaluation and monitoring may provide a more cost effective way of assessing policy effectiveness.

Recommendation 17: In some circumstances it may be advisable to focus on specific outcomes when data availability in administrative data sets give some chance of constructing a reasonable control group.

Recommendation 18: In some circumstances it may be advisable to focus on specific policy features to at least allow identification of impacts for specific groups of recipients.

Recommendation 19: To obtain high quality impact evaluations, departments need to be prepared to consider evaluation issues at the time of policy design, in particular with a view to embedding aspects of randomisation into the programme delivery.

Recommendation 20: Consideration should be given to the establishment of an independent body responsible for 1) peer review of central government department evaluations before, during and after the commissioning process – possibly using a peer review college of experts; 2) long term evaluation of major policy initiatives.

Recommendation 21: To facilitate long-term impact evaluations, government should establish better protocols for confidential sharing of administrative data with trusted researchers.

6. Conclusions

Our review of evaluations suggests that the quality of cost-effectiveness reports varies widely. Based on the criteria we have considered, we found evidence of high quality evaluations in the areas of active labour markets and education. In contrast, evaluations in the areas of business support and spatial policy were considerably weaker. Using the five-part scale described in box 1 we ranked six of the business support evaluations as level 2, and the final one as level 1. If anything, reports in the area of spatial analysis did slightly worse with three ranking level 1, and the remaining 4 ranking level 2 (we couldn't rank one report). The available reports generally provided very little technical detail. Regardless, the approaches adopted were not sufficiently robust to give us any confidence in the estimated impacts (or, as a result, in the cost-effectiveness evaluations). At best, for a few reports, a defensible approach was adopted but implementation weak (or impossible to assess on the basis of information provided).

The evaluations of active labour markets and education policies are far better in this regard. In the area of education we ranked five reports at level 4 and one at level 5. Three reports ranked level 2, while only one ranked level 1. The labour market reports had a similar profile (one ranked at level 5, four ranked at level 4, three at level 3, one at level 1, with one difficult to grade on the basis of the preliminary report. Perhaps unsurprisingly, the active labour market evaluations were arguably the strongest in terms of establishing more credible impacts and costs-effectiveness evaluations. But many of the education evaluations were also of high quality, despite the fact that they face difficulties that are not dissimilar to some of those faced in the evaluation of spatial and business support. Even in these two areas, however, there were a

couple very weak reports. Our detailed discussion above also highlights the fact that the stronger reports could still be improved along a number of dimensions.

Ranking reports in terms of their overall quality is not an exact science, but the marked differences between education and labour markets on the one hand, and spatial and business support on the other, should be clearly demonstrated. If we take level 3 on the Maryland scale as the minimum necessary for having any confidence that the impacts detected may be attributed to policy, then our overall assessment would be that none of the business support or spatial policy evaluations provided convincing evidence of policy impacts. In contrast, 6 out of 9 of the education reports and at least 6 (possibly 7) out of 10 labour market reports were of sufficient standard to have some confidence that the impacts could be attributed to policy.

For business support and spatial policy, there appears to have been an over-reliance on self-reported additionality and on poorly explained and poorly justified approaches to 'correct' for selection in to treatment. We recognise that these are areas where evaluation is, arguably, more difficult, but the gulf between best practice and the evaluations cannot be attributed to this alone. Indeed, in some situations, the structure of the programme and the data collected for the evaluation would have allowed for careful impact evaluation, but this did not happen. In other situations, use of available administrative data and better methodologies could have provided far more convincing data.

We were asked to identify low cost ways to improve cost-effectiveness evaluations. To the extent that issues we identify are about moving closer to best practice they would fit this criteria. Other aspects we have highlighted are likely to prove more difficult to address. The first of these is the need to change practice in areas of policy that are currently very poorly evaluated. Our report has identified two of these – but there will be other areas across government. The second difficult cross-cutting issue is the need to be realistic about what evaluation can achieve, to better focus evaluations, and to think how to trade-off the scope of evaluations against the robustness of the results. At the moment our review of evaluations suggests that the balance is arguably tipped to far in favour of large scale evaluations that fail to establish the cost-effectiveness of interventions. Third, even the best methodology may struggle to identify the impact of policy if evaluation is not embedded from the earliest stages of policy design. If we want robust evaluation of the cost-effectiveness of very expensive policies there may need to be

much more realism about the need to pilot or find other ways to provide robust evaluation. Finally, there is the fundamental issue of how policies are evaluated and by whom. Our current system favours early evaluations undertaken by government departments that have large vested interests. An alternative system would see far more independent evaluation, over longer time periods. Such a system would need to be able to embed policy understanding (that sits in departments) in to the evaluation process as well as ensuring that 'ownership' of the evaluations (and hence the need to act on poor cost-effectiveness outcomes) was not reduced by a move to more independence. These issues are complex, but they do need addressing if we wish to produce cost-effectiveness evidence that is fit for purpose.

Appendix 1: Final lists of evaluation projects for retrospective review

Education

Every Child Counts:

https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-RBX-10-07

Every Child a Reader

https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-RR114

Achievement for All:

https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-RB176

National Evaluation of Sure Start local programmes:

https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-RR073

Social and Emotional Impact of Learning (SEAL) programme in Secondary Schools

https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-RB049

Key Stage 2 career-related learning pathfinder evaluation
https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-RB116

Evaluation of the Education Maintenance Allowance Pilots

https://www.education.gov.uk/publications/standard/publicationDetail/Page1/RR678

Activity Agreement Pilots – Quantitative Evaluation

https://www.education.gov.uk/publications/standard/publicationDetail/Page1/DCSF-RR096

Evaluation of the National Citizen Service pilots, recently published (May) on half of both DfE and the Cabinet Office

http://www.natcen.ac.uk/media/898405/ncs-evaluation-interim-report.pdf

Teens and Toddlers

https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR211.pdf

Active Labour Markets

Evaluation of the Job Outcome Target Pilots: DWP in house quantitative study

http://research.dwp.gov.uk/asd/asd5/rports2005-2006/rrep316.asp

Employment Retention and Advancement demonstration (ERA). Final evidence report containing cost benefit analysis

http://research.dwp.gov.uk/asd/asd5/rports2011-2012/rrep765.pdf

Evaluation of the New Deal for Disabled People: Impacts and cost-benefit analyses

http://research.dwp.gov.uk/asd/asd5/rports2007-2008/rrep430.pdf

The econometric evaluation of New Deal for Lone Parents

http://research.dwp.gov.uk/asd/asd5/rports2005-2006/rrep356.pdf

http://statistics.dwp.gov.uk/asd/asd5/working_age/wa2003/wae147rep.pdf

Evaluation of the Fair Cities Pilots 2007; qualitative study that aims to provide guidance on cost-effectiveness

http://research.dwp.gov.uk/asd/asd5/rports2007-2008/rrep495.pdf

Pathways to Work for new and repeat incapacity benefits claimants: Evaluation synthesis report

http://research.dwp.gov.uk/asd/asd5/rports2007-2008/rrep525.pdf

The introduction of Jobcentre Plus: An evaluation of labour market impacts

http://research.dwp.gov.uk/asd/asd5/rports2011-2012/rrep781.pdf

Gateway to Work New Deal 25 Plus pilots evaluation

http://research.dwp.gov.uk/asd/asd5/rports2005-2006/rrep366.pdf

Work Based Learning for Adults

http://research.dwp.gov.uk/asd/asd5/rports2005-2006/rrep390.pdf

http://statistics.dwp.gov.uk/asd/asd5/working_age/wa2004/187rep.pdf

Early Impacts of the European Social Fund 2007-13

http://research.dwp.gov.uk/asd/asd5/ih2011-2012/ihr3.pdf

Business Support

Evaluation of Regional Selective Assistance 1991-1995
http://www.bis.gov.uk/files/file22008.pdf

Evaluation of Regional Selective Assistance (RSA) and its successor, Selective Finance for Investment in England (SFIE)
http://www.bis.gov.uk/files/file45548.pdf

Evaluation of Grant for Research and Development and Smart 2009
http://www.bis.gov.uk/files/file52026.pdf

Evaluation of Smart (including SPUR) 2001: Final Report
http://www.bis.gov.uk/files/file22000.pdf

Economic Impact Study of Business Link Local Service
http://www.bis.gov.uk/files/file40289.doc

Economic evaluation of the small firms loan guarantee
http://www.bis.gov.uk/files/file54112.doc

Evaluation of the Manufacturing Advisory Service: Main Report
http://www.berr.gov.uk/files/file38877.pdf

Spatial Policy

National Evaluation of the Local Enterprise Growth Initiative Programme - Final report
http://www.communities.gov.uk/publications/regeneration/lgipfinalreport

Evaluation of the Mixed Communities Initiative: Demonstration Projects - Final report
http://www.communities.gov.uk/publications/housing/mixedcommunitiesinitiative

Regenerating the English Coalfields - Interim evaluation of Coalfields Regeneration Programmes
http://www.communities.gov.uk/documents/regeneration/pdf/324761.pdf

Evaluation of the National Strategy for Neighbourhood Renewal - Final report
http://www.communities.gov.uk/publications/communities/evaluationnationalstrategy

Evaluation of the National Strategy for Neighbourhood Renewal: Econometric modelling of neighbourhood change
http://www.communities.gov.uk/publications/communities/evaluationnationalchange

CLG (2009) Evaluation of the National Strategy for Neighbourhood Renewal: Improving educational attainment in deprived areas.
http://www.communities.gov.uk/documents/communities/pdf/1490497.pdf

BERR (2009) Impact of RDA spending – National Report – Volume 1 – Main report
http://www.berr.gov.uk/files/file50735.pdf

BERR (2009) Impact of RDA spending – National Report – Volume 2 – Regional Annexes
http://www.bis.gov.uk/files/file50736.pdf

The Single Regeneration Budget – Final Report
http://www.landecon.cam.ac.uk/staff/publications/ptyler/SRB_part1_finaleval_feb07.pdf
http://www.landecon.cam.ac.uk/staff/publications/ptyler/SRB_part2_finaleval_feb07.pdf
http://www.landecon.cam.ac.uk/staff/publications/ptyler/SRB_part3_finaleval_feb07.pdf

7. About the authors

Henry Overman is Professor in Economic Geography in the department of Geography and Environment at the London School of Economics and since April 2008, the director of the Spatial Economics Research Centre. His current research interests include the causes and consequences of spatial disparities and the impact of urban and regional policy. His research has been published in leading economics journals (The Review of Economics Studies and The Quarterly Journal of Economics) and leading economic geography journals (Environment and Planning and Journal of Economic Geography). He continues to publish in journals from both disciplines and, in August 2007, took over the joint-editorship of the inter-disciplinary Journal of Economic Geography. He has extensive experience in both critiquing and undertaking evaluation conducted by government and other organisations as part of academic and consultancy projects. He has undertaken academic research evaluating the impact of Regional Selective Assistance, the Single Regeneration Budget the Local Enterprise Growth Initiative. In addition, he has conducted a study for BIS critiquing their existing evaluations and examining the scope for the wider use of programme evaluation techniques in BIS commissioned evaluations. Together with Steve Gibbons he undertook a DfT project critiquing the methods used to evaluate the impact of transport infrastructure and piloted the use of programme evaluation techniques. He continues to work on the application of these techniques as part of an ESRC funded project on ex-post evaluation of transport investments. He has provided policy advice to, amongst others, the European Commission, Department for International Development, Department for Business Enterprise and Regulatory Reform, Department for Communities and Local Government and the Department for Transport. He is also affiliated with the Centre for Economic Policy Research' s International Trade Programme.

Steve Gibbons is Reader in Economic Geography at the LSE. He is also Research Director of the Spatial Economics Research Centre at the LSE and a Research Associate of the Centre for Economic Performance and the Centre for Economics of Education. His research applies the econometric techniques of programme evaluation to research questions in the economics of education, transport, health, housing markets and crime, and he is an expert on these methods. He has worked on evaluations of the effects of policy related to choice and competition policy in health care (Cooper at al 2011, 2012), road and rail transport infrastructure improvements (Gibbons and Machin 2005, Gibbons et al 2010), and housing benefit (Gibbons and Manning

2006). Many of these studies involve rigorous critiquing of existing studies (including those commissioned by government) and improvements on those studies through the application of difference-in-difference analysis appropriate to policy evaluation work and include elements of cost-benefit analysis. He has published on these issues in leading economics and urban journals including the Economic Journal, Journal of the European Economic Association, Journal of Labor Economics, Journal of Urban Economics, Journal of Public Economics and Urban Studies. He has carried out consultancy work for DfT on evaluating the effects of transport infrastructure projects on firm productivity, employment and housing values, and for the Eddington study, on linkages between transport and the labour market. He has been member of the Home Office Criminal Justice Steering Group and the Millennium Cohort Survey 4 Advisory Group. He has undertaken funded research on poverty for the Joseph Rowntree Foundation, Widening Participation in Higher Education for the ESRC, and on the evaluation of transport infrastructure improvements.

Sandra McNally is Director of the Education and Skills Programme at the Centre for Economic Performance, London School of Economics. She is also Professor of Economics at University of Surrey. She has vast experience of conducting evaluation in the economics of education and of reviewing evidence about 'what works' in this area. She has extensive experience in critiquing existing studies and implementing ex-post evaluation studies and applying methods that aim to identify the causal impact of interventions. She has participated in government commissioned evaluation of Excellence in Cities, Excellence in Primary Schools, Aimhigher, and Pupil Learning Credits. In addition she has undertaken projects facilitated, but not commissioned by, government (e.g. by provision of data). These include past evaluations of the pilot of the National Literacy Strategy (the literacy hour) and a current evaluation of the pilot of the synthetic phonics programme implemented as a result of the Rose Review in 2005/06. She also has experience of devising both an intervention and its evaluation. This is in the context of a field experiment in secondary schools in London where the objective to ascertain the effects of providing careers related information to pupils (work-in-progress) The Education and Skills Programme of CEP has recently been appointed as one of ten evaluators of the panel for the Education Endowment Fund. A project has recently been awarded where CEP will evaluate a pilot aimed at closing the gap between disadvantaged pupils and their peers in the context of a randomized control trial. She has previously advised the Financial Services Authority on evaluation methods (McNally and Meghir, 2009) and has written several review articles or book

chapters where evaluation studies or methods are discussed (e.g. Machin and McNally, 2012; Emmerson, Meghir, McNally 2005). All this work takes account of international studies about evaluation and includes comparisons with related studies in other countries.

Appendix: Evaluations in the area of Education policy

This appendix provides details of the evaluations considered in the area of education policy. The structure of the template was agreed following discussions with the National Audit Office. In completing the templates, for reasons of both feasibility and presentation, we have made use of source material from the original evaluations without any attempt to provide detailed attribution (e.g. through the use of quotes, or the provision of page numbers).

Achievement for All: National Evaluation

Policy objectives

'Achievement for All' (AfA) was conceptualised as a means to support schools and Local Authorities to provide better opportunities for learners with special educational needs and disabilities (SEND) to fulfil their potential. There were three main strands: (1) assessment, tracking and intervention; (2) structured conversations with parents; (3) provision for developing wider outcomes (attendance, behaviour, bullying, developing positive relationships). This was developed as a Pilot in selected schools within ten LAs.

Scope of evaluation

- To examine the impact of AfA on a variety of outcomes for children and young people with special educational needs and disabilities (SEND).
- To find out what processes and practices in schools were most effective in improving these outcomes.

Overall methodology

- Surveys of teachers and parents in relation to outcomes for Strand 2 (structured conversations with parents) and Strand 3 (provision for developing wider outcomes). Online surveys of teachers and parents of children and young people with SEN in participating schools and some comparison schools.
  - Surveys conducted at three points in time: Jan 2010, Jan 2011, June 2011.
  - For the teachers survey, final sample of 4,794 teachers in AfA schools and 196 teachers in comparison schools.
  - For the parent survey, 294 parents in AfA schools and 13 parents in comparison schools.
- Attendance data provided by participating LA. This was used to calculate the percentage attendance for each pupil in the target cohort in the year prior to the AfA pilot (2008/09) and during the two years of the pilot (2009/10 and 2010/11), which was used to examine changes in attendance patterns. The final sample was 8,656 pupils attending AfA schools and 194 attending comparison schools.

- Academic attainment data provided by the National Strategies at three time points: December 2009; December 2010 and July 2011. To assess relative academic progress of pupils in the sample, draws upon national statistics supplied by the DfE. Compares changes in Maths and English scores for pupils in AfA schools to pupils with and without SEND nationally in England.
- Collect school-level data from administrative data and an online school survey to look at the way school-level contextual and compositional features and AfA implementation processes and practices impact upon progress on pupil-level outcomes.
- The qualitative component consists of interviews with local and regional lead professions; school case studies of 20 AfA schools (5 visits per school), pupil case profiles for 87 pupils across case study schools; informal data collection at a range of events (e.g. 10 launch conferences).

Impact evaluation

- To look at the impact of AfA on pupils' academic progress in English and Maths, measures of pupil progress in AfA schools are compared to an estimate of average progress made by pupils nationally (both those with SEND and without SEND).
- A multi-level analysis is used to examine the characteristics of pupils and schools that are associated with pupils' academic progress. This is used, for example, to show that pupil progress is associated with secondary schools that 'show greater fidelity to the structured conversation model'.
- The impact of AfA on parental engagement and confidence is assessed by using AfA schools only as there were insufficient returns from parents in comparison schools. This is based on a sample size of 283 parents.
- The impact of AfA on 'positive relationships' of pupils with SEND is considered by comparing those attending AfA schools (N=4,562) with those attending comparison schools (N=193).
- The impact of AfA on attendance is analysed by comparing pupils attending AfA schools (N=9,115) with those attending comparison schools (N=223) using data extracted from LA records.

Policy details

The Achievement for All pilot involved ten LAs selected by the Department for Children, Families and Schools (now DfE). Each LA selected schools to participate and in total there were 454 schools (including primary and secondary mainstream schools, special schools and a small number of pupil referral units) over a two-year period.

No information is provided about how LAs or schools were selected. The initiative was introduced in 2009 and the evaluation report was published in November 2011.

Data

The quantitative analysis uses surveys of teacher and parents (see above). It also uses data on pupil outcomes provided by the National Strategies and the DfE.

The authors estimate average pupil progress by using Teacher Assessment in different year groups over a 19 month period. This involves combining Teacher Assessments across different year groups (1, 5, 7, and 10) and converting them to a common scale. The details of exactly how and when the Teacher Assessments were conducted are vague. The estimates of progress from national data are based on Key Stage Assessments.

Costs

The AfA Pilot received £31 million over a two year period. No further information is provided (e.g. on expenditure by type of LA).

Outcome variables

Outcome variables included measures of progress in English and Maths (described under 'data' above), measures of behaviour, attendance and positive relationships.

Control group

For some parts of the analysis, outcomes are compared to pupil outcomes nationally using administrative data. The comparisons are made either with pupils classified as having special educational needs and disabilities or all pupils.

For other parts of the analysis, there is either no control group or a very small control group. No information is provided on how the control group was selected or their comparability with the treatment group.

Methodology details

Progress in English and Maths was estimated for AfA schools over a 19 month period – between December 2009 and July 2011. This was compared to an estimate of average progress made by other pupils with SEND nationally over an equivalent period of time.

In other analyses about progress in English and Maths, various school and pupil attributes were included in a multi-level regression model. This was used to see how included variables were associated with progress. However, the link with AfA was not clear in this analysis. For example, with regard to progress in English, students with particular categories of special needs were found to make greater progress and those with other categories were found to make less progress. However, this analysis is uninformative about the effects of the AfA pilot. However, in other cases, there was a link with the AfA programme: for example, it was found the pupil progress was greater in Maths in those secondary schools that involved parents more often in reviewing academic targets etc. This sort of analysis was used for other types of outcome (e.g. parental engagement and confidence). The analysis tries to associate variables linked with good implementation of the AfA to these outcomes. Most schools are AfA schools and thus the variation is coming from the extent to which schools implement particular practices rather than being in the Pilot.

Comparisons between the treatment and control group are used for changes in "pupils' positive relationships" (as reported in the teacher survey). Graphs are shown of the change in treatment schools (N=4562) to the change in control schools (N=193). This is also used to look at attendance. The type of 'multi-level regression analysis' described above is also used to look at these outcomes.

Internal validity

The methodology used in this study is not appropriate for an impact study as it compares treatment schools to national averages, without regard for the fact that the school context will be different in treatment schools and schools nationally. Furthermore, basic details are not provided such as the characteristics of schools selected to be part of AfA, why they were chosen, and how representative they are of schools nationally.

Inference

Tables in the text show coefficients and statistical significance. Guidance is also given about how to interpret coefficients. Detailed regression tables are provided in an appendix.

External validity

This is not discussed.

Cost effectiveness

The authors interpret their results in a causal way. For example, they state that the AfA had a significant impact upon progress in English and Maths among pupils with SEND. They say that effect sizes range from small to very large but in all cases big enough to be practically meaningful ('for instance, pupils in Year 10 were on course to achieve a greater number of A*-C GCSEs'). They state that 'the AfA pilot proved to be very successful in narrowing the well established achievement gap between pupils with and without SEND'.

There is no discussion about cost-effectiveness.

Overall assessment

This analysis does not follow what would be considered good practice in the programme evaluation literature. Overall, it would rate at level 1 on the Maryland Scale (possibly level 2 if willing to view the 'national average' as providing a 'comparison group' – albeit and invalid one). The data collected and qualitative assessment is of some use in understanding how schools implemented AfA and perceptions of what worked well etc. However, the evaluation does not give credible quantitative findings on impact.

International comparators

Evaluation of programmes to help special needs children include the following:
Hanushek, Kain and Rivkin assess effects by looking changes over time for students who move in and out of targeted programmes (controlling for endogeneity bias in various ways).
Hanushek, E.A., J.F.Kain and S.G. Rivkin. (2002). Inferring Program Effects For Special Populations: Does Special Education Raise Achievement for Students with Disabilities? *Review of Economics and Statistics*, 84(4), 584-599.

Keslair, Maurin and McNally, (2011) use differences across school context in the probability of being assigned to a special needs programme to assess the impact on attainment in primary school (in England). http://cee.lse.ac.uk/ceedps/ceedp129.pdf

Documents examined

Achievement for All: National Evaluation: Final Report. Neil Humphrey and Garry Squire. Research Report DFE-RR176. November 2011.

https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR176.pdf

Activity Agreement Pilots – quantitative evaluation

Policy objectives

The Activity Agreement Pilot is an initiative aimed at testing the effectiveness of conditional financial incentives along with intensive support and brokerage of tailored activities in re-engaging young people aged 16-17 who had been NEET for at least 20 weeks.

An Activity Agreement (AA) is a personally negotiated contract between a Connexions Personal Advisor and the young person. It is an individually tailored and agreed programme of activities designed to break down barriers to participation and identifies specific steps that the young person will take to move into education, employment or training. Whilst participating, young people receive one-to-one support and advice and a weekly allowance – paid only if the young person fulfilled their weekly agreement.

Scope of evaluation

- A quantitative evaluation, using surveys of young people to measure the impact of the pilots in comparison to a number of control areas.
- A programme theory element, focusing on testing some aspects of the policy to identify what works, what does not and the reasons for this.
- A process evaluation, examining the ways in which the pilots have been set up and delivered and the main issues associated with their implementation.

Overall methodology

- Description of characteristics of participants and non-participants.
- Analysis of participants' experiences of AAs from wave 1 interviews with participants, parallel interviews with parents of some participants and wave 2 follow-up interviews with a sub-sample of participants.

- Analysis of the impact of AAs on participants by comparing them to those in a control group. The full (unmatched) sample consists of 3,331 interviews in pilot areas and 2,291 in comparison areas.

Impact evaluation

- Treatment group consists of those taking up an Activity Agreement. They are matched to those in a comparison group (in LAs not part of the Pilot) using propensity score matching.
- Surveys of those in the treatment and comparison group, including self-reported educational and employment outcomes, attitudes towards learning and work. The surveys used a mixture of face-to-face and telephone interviewing.
- Relatively short-term outcomes – outcomes for participants only captured for a period of one year after first becoming NEET and 32 weeks after becoming eligible for AA.

Policy details

Eight pilot areas were selected, implementing one of three variants of the pilot, which differed in the level of the weekly payment available to the young person and in one variant a payment to the parent. The pilots began in April 2006 and initially ran for two years. Survey interviews for this evaluation were carried out between January 2007 and March 2008.

Data

Participants in treatment and control areas are asked about participation in a range of employment related activities within 12 months of becoming NEET (with details on these activities). Average values are reported in an appendix. There are also variables used for the matching which are tabulated in an Appendix. These variables might come from the surveys or from administrative data provided by Connexions (not clear).

Costs

The 2005 budget allocated £60 million to this pilot. There is a description of the payments per person in each variant of the scheme (£20 per person to the young person; £30 per person to the young person; £20 per week to the young person and £30 per week to their parent). However, there is no formal analysis of costs (or cost-effectiveness analysis) in the report.

Outcome variables

The key measures of impact are based on the self-reported activity status of the young person: involvement in paid work, training or education activities.

Control group

The control group was drawn from areas that were not taking part in the Pilot. The sample records were provided by Connexions in each pilot and comparison area. From these records, the researchers were able to find 2,291 participants in comparison areas (and 3,331 participants in treatment areas). It is not clear why the eight pilot areas were selected for the AAs and the extent to which these areas are comparable to areas used for the control group. However, the propensity score matching is clearly explained. Survey participant characteristics across treatment and control areas are compared before and after matching. This is shown in a detailed table (which unfortunately does not report sample size in each case). However, the matching does a good job in making the treatment and comparison areas more similar across a range of characteristics and this is very clearly shown.

Methodology details

Eligible participants who chose to take up an Activity Agreement are compared to non-eligible participants (who live in areas not covered by the Pilot). These groups are matched using propensity score matching, such that observable characteristics are similar. However, the risk of self-selection bias is acknowledged. The report contains an appendix where the issue of using eligible participants versus eligible participants who chose to take up AA agreements is explained. The authors choose to use eligible participants as the relevant group because of low take-up of AAs in treatment areas (estimated at about 20%).

Internal validity

The fact that the authors choose not to report the 'intention to treat' effect damages the credibility of this analysis. The 'intention to treat' effect (i.e. estimated on the eligible population versus the control group) is of primary interest in the programme evaluation literature. The authors discuss this issue in some detail in the appendix and give an estimate of the 'intention to treat' effect for one outcome (i.e. personal development activities). However, they do not say what the 'intention to treat' effects would be on the outcomes of primary interest in this analysis (self-reported activity status of young people). Instead they focus on the effect as estimated for eligible participants versus the control group. If there is positive selection bias (i.e. those who took-up the offer of an AA were more likely to return to work/education even without the programme, relative to those in the control group), then the estimates will all have an upward bias. *Inference*

The outcome variables for those taking up the AA agreement and their matched counterparts (in the control group areas) are set out in a table. The difference is shown, together with a star to indicate significance at the 10 percent level. Standard errors are not reported. The text describes effects as 'small'. However, this is not always accurate as the estimates need to be interpreted in the context of average values among those in the comparison group. When viewed in this context, the 'effects' are fairly large (although for the reasons discussed above, these estimates could well be biased).

External validity

These estimates do not have external validity because they are comparing self-selected participants in an AA agreement to those in a control group. Furthermore, the treatment areas are in 8 pilot regions. We do not know the basis on which these have been selected for the Pilot.

Cost effectiveness

The authors of their report interpret their findings as suggesting that the AA participation had a large impact on participating in personal development activities, but beyond that suggest that effects were modest. There is no attempt to compare benefits to costs.

Overall assessment

The use of propensity-score matching on treatment and comparison groups that emerge from the policy pilot designplaces this report at level 4 on the Maryland scale in terms of overall research design, although there are a number of weaknesses in write up and implementation.

For this report, we would particularly highlight the issue of external validity. It is bad practice not to report 'intention to treat' effects. Even in the absence of selection bias into the treatment (as is likely to be the case), the take-up of a policy should be part of any analysis about whether or not it was effective. The 'intention to treat' effect is of great interest for policy makers and it is very disappointing that this is not reported.

However, it is also of interest to try to scale up results for participants (impact of 'treatment on the treated'). The researchers do this in an appendix. For one outcome variable (personal development activities), the report gives the 'intention to treat' effect and then scale up the result to account for the fact that the participation rate in AA agreements was only 20% of the eligible population. The report also gives an alternative estimate based on directly comparing participants to those in the comparison group. In this case, the report shows that the two estimates (for the 'effect of treatment on the treated') is similar. For other outcomes, the researchers comment that they found some variability between the two approaches but similar effects overall. For the sake of transparency, the estimates should have been compared for the outcomes of primary interest in this analysis using both methods.

The short-term nature of this evaluation is another strong limitation.

International comparators

This programme has similarities to the Educational Maintenance Allowance. See overview of this programme and references to international comparators. However, this programme is aimed at those who have dropped out of education. There is a lot of evaluation evidence in the US for programmes to help high school drop-outs (and plenty of RCTs). However, it appears that many programmes are ineffective and the ones that work can be quite costly. See the discussion in Heckman, J.J., and L. Lochner, (2000), "Rethinking Education and Training Policy: Understanding the Sources of Skill Formation in a Modern Economy," in S. Danziger and J.

Walfogel (eds), *Securing the Future: Investing in Children from Birth to College*, Russell Sage Foundation: New York.

Documents examined

Activity Agreement Pilots – Quantitative Evaluation. Emily Tanner, Susan Purdan, John D'Souza and Steven Finch. National Centre for Social Research. DCSF – RR096. April 2009

https://www.education.gov.uk/publications/eOrderingDownload/DCSF-RR096.pdf

Evaluation of Education Maintenance Allowance Pilots: Young People Aged 16 to 19 Years: longitudinal quantitative evaluation

Policy objectives

The EMA pilots were introduced to assess whether offering a monetary allowance to young people from low income families would encourage them to remain in education after the end of compulsory education. The policy context was a slowing down in the rate of participation in post-16 education. There had been an increase in the 1980s and early 1990s but then remained at about the 1994 level (just over 70%). In particular, there were concerns about the male-female gender gap (7% higher for females) and the socio-economic divide. There were also concerns about retention as the participation rate drops dramatically with age (69.7% for 16 year olds; 57.7% for 17 year olds and 37.1% for 18 year olds in 2000).

Scope of evaluation

- The EMA is one of the most extensive evaluations of any initiative that the Department for Education has ever commissioned. The statistical evaluation is considered here (the final of four reports). This is one element of a larger exercise involving a range of research methods. 12 reports are listed in the Appendix.
- The longitudinal quantitative evaluation involved large samples of young people who finished compulsory education in the summers of 1999 and 2000.
- The aim of the evaluation was to estimate the impact of EMA on participation, retention and achievement in post-16 education.
- The evaluation shows detailed findings on these participation and retention for young people up to the age of 18; then to the age of 19; then the qualifications that young people achieved over 3 years following the end of compulsory education.

Overall methodology

- Longitudinal cohort study involving large surveys of random samples of young people in ten of the original 15 EMA pilot areas and 11 control areas.
- Eight datasets produced from four interviews with two cohorts of young people (and their parents at Wave 1) conducted at annual intervals.

- Weights were designed to correct for potential sources of bias arising from restrictions on the sampling procedure and from possible differences in initial non-response so that results could be produced that were representative of all young people in the pilot and control areas.
- Dual approach to analysis, using both descriptive and 'matching' techniques. Descriptive techniques seen as complementary to matching because it allows data to be explored at a high level of disaggregation. Also, data can be weighted to account for attrition. However, it cannot provide a measure of impact.

Impact evaluation

- 10 Pilot areas and 11 control areas selected.
- Matching at Local Authority level to compare similar areas in terms of Pilot and control.
- Propensity Score Matching (PSM) at individual level to achieve a control group where each individual is as alike to their counterpart in the pilot area as possible using observed characteristics.
- Outcomes considered are participation, retention and qualifications attained between the ages of 16 and 19.

Policy details
The EMA is an allowance paid to 16-19 year olds (or in some areas to their parents), eligibility for which is dependent on parental income. The pilot provision started in September 1999 in 15 Pilot areas. It was decided to roll out the policy nationally in 2002 and this had taken place by September 2004. The Coalition Government announced that this programme would be discontinued in 2011.
Data
Longitudinal cohort survey involving large random sample surveys of young people and their parents in ten EMA pilot areas and 11 control areas. Two cohorts selected – the first left compulsory education in the summer of 1999 and first interviewed between November 2000 and April 2001. The second left compulsory education in the summer of 2000 and were first interviewed between November 2000 and April 2001. They were interviewed 3 years later (43% of the original sample).
Costs
The report explains the structure of EMA and the amounts given to eligible individuals in each variant of the Pilot and also the national scheme. However, the overall costs of the Pilot and national scheme are not discussed here.
Outcome variables
Outcome variables are participation in full-time education, retention in full-time education and achievement (qualifications) between the ages of 16 to 19.
Control group
There were two main stages to finding a control group. In the first stage, pilot areas were matched to potential control areas. In the second stage, individuals in treatment areas are matched to those in control areas. The process is very thoroughly explained in the first report.

Methodology details
Comparison of treatment and control group using propensity score matching. Also, a descriptive approach that uses the full sample – also comparing treatment and control. This applies weights to account for attrition and is also regard as a useful check on the direction of the findings from the PSM approach. Full details provided.
Internal validity
The detailed and careful analysis is convincing on the effects of the intervention with regard to participation and retention in the first three waves of the study. Positive effects are found on participation at age 16 of 5.9 percentage points. This comes both from work and training (-3.4 percentage points) and the NEET group (-2.4 percentage points). The effects are stronger for young men than young women. Similar effects are found for retention at age 17. At age 18, effects are upheld for men but not statistically significant for women.
No effect is found on post-16 qualification attainment. The report is not confident about the robustness of this finding because of a high rate of attrition and inconsistencies between administrative data and young people's self-reports (affecting 15% of the sample). Furthermore, the report suggests that the set of variables used in the PSM matching procedure, whilst suitable for modelling participation and retention, might not have been suitable for examining achievement.
Inference
Mostly explained in detail. However, standard errors are not provided in the tables.
External validity
To some extent, this is provided through a comparison between the descriptive results (which use the full sample and weighting) and the matching analysis. There is also analysis by different subgroups. The Executive Summary gives estimates of the national impact of EMA (beyond the Pilot). However, this is not a focus of the report. It would have been helpful to have a section which discussed this explicitly.
Cost effectiveness
There is no cost-effectiveness analysis in the main impact report.
However, the authors produce a back-of-envelope cost-benefit analysis in published work on some aspects of the programme.
"The EMA increased the percentage of individuals from income-eligible families completing two years of post-compulsory education by 6.7 percentage points, from 54.3 percent to 61.0 percent. In the first year (second year), one-third (two-thirds) of this increase was from individuals who would otherwise have been in paid employment. This means that those brought into education would need to experience a real increase in future earnings of 6.2 percent as a result of the additional two years of education for the program to break even, allowing for the opportunity cost of education. Allowing £3,000 for the extra annual cost of educating those who stay on in secondary education increases the required return to education for the two years to 7.7 percent. Research into the returns from staying on in postcompulsory education suggests that the returns are in fact 11 percent for males and 18 percent for females. There may well be other benefits of the policy: the government might value the redistribution to lower-income families with children; infra-marginal individuals may reduce hours of work and increase effort put into education; there may be crime reductions."[5]

[5] http://www.ifs.org.uk/wps/wp0511.pdf

Overall assessment

This is a very careful evaluation. The approach adopted would rate 4 on the Maryland scale. The treatment and control groups are very carefully matched and the control group remained uncontaminated by subsequent policy to roll-out the programme until the fieldwork had been completed.

The early evaluation showed that EMA significantly increased participation and retention. This is likely to have influenced the Government to roll-out the policy nationally in 2002 (completed by September 2004). The Coalition Government recently decided to cease this programme (in 2011)

The evaluation does not provide robust results on post-16 achievement. This has to do with three factors: (a) a high rate of attrition; (b) a high degree of mismatch (15%) between young people's self reports of qualifications and administrative data – which mainly affects the first cohort; (c) variables appropriate for matching with regard to participation/retention are not necessarily appropriate with regard to educational attainment.

A way of overcoming the matching problems would have been to randomise areas into the treatment and control group. However, it is difficult to know how to avoid the problems of attrition except to provide a link between the original participants and their qualification through an identifier which would allow linkage through administrative data sets (now technically possible). Alternatively, a new random sample of individuals across treatment and control areas could have been taken for a separate study about achievement (although this would have been costly).

International comparators

A number of countries have introduced means-tested conditional grants in an attempt to encourage students to stay in school. Examples include PROGRESA in Mexico and Familias en Acción in Colombia. They have been evaluated (respectively) in a RCT and difference-in-differences framework.

- Attanasio, Orazio, Emla Fitzsimons, Ana Gomez, Diana Lopez, Costas Meghir and Alice

Mesnard. 2006. "Child education and work choices in the presence of a conditional cash transfer programme in rural Colombia." Working Paper W06/13. London: Institute for Fiscal Studies.

- Attanasio, Orazio, Costas Meghir, and Ana Santiago. 2007. "Education Choices in Mexico:

Using a Structural Model and a Randomised Experiment to Evaluate Progresa." Working Paper EWP05/01. London: Institute for Fiscal Studies.

Other related papers include Dynarski (2003), who examines the impact of incentives for college and Angrist, Lang, and Oreopoulos (2006), who use a randomised trial at a Canadian university to examine the impact of increased financial incentives, increased non-financial support, and both increased financial and non-financial support.

- Dynarski, Susan. 2003. "Does Aid Matter? Measuring the Effect of Student Aid on College

Attendance and Completion." *American Economic Review* 93(1): 279–88.

- Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2006. “Lead Them to Water and Pay

Them to Drink: An Experiment with Services and Incentives for College Achievement.” Working Paper 12790. Cambridge, MA: National Bureau of Economic

Documents examined

Evaluation of Education Maintenance Allowance Pilots: Young People Aged 16 to 19 Years Final Report of the Quantitative Evaluation
Centre for Research in Social Policy: Sue Middleton, Kim Perren, Sue Maguire, Joanne Rennison
Institute for Fiscal Studies: Erich Battistin, Carl Emmerson, Emla Fitzsimons. Report to Department of Education and Skills RR678
https://www.education.gov.uk/publications/eOrderingDownload/RR678.pdf

Education Maintenance Allowance: The First Year, A Quantitative Evaluation
Centre for Research in Social Policy: Karl Ashworth, Jay Hardman, Woon-Chia Liu, Sue Maquire and Sue Middleton
Institute for Fiscal Studies: Lorraine Dearden, Carl Emmerson, Christine Frayne, Alissa Goodman, Hidehiko Ichimura and Costas Meghir. Report to Department of Education and Employment RR257
http://eprints.ucl.ac.uk/18495/1/18495.pdf

Education Subsidies and School Drop-Out Rates. Lorraine Dearden, Carl Emmerson, Christine Frayne
Costas Meghir. The Institute for Fiscal Studies. WP05/11
http://www.ifs.org.uk/wps/wp0511.pdf

Every Child a Reader

Policy objectives

Every Child a Reader (ECaR) offers a layered, three-wave approach to supporting children with reading in Key Stage 1. Wave 1 is ‘quality first teaching’ aimed at all children through class based teaching. Wave 2 is a intervention aimed at groups of children (or potentially one-to-one) who can be expected to catch up with their peers with some additional support. Wave 3 offers intensive reading support in the form of a one-to-one programme for children who have been identified as having specific support needs. The main intervention under Wave 3 is ‘Reading Recovery’, an intensive programme lasting approximately 20 weeks for the lowest attaining 5 per cent of children aged five or six. The ECaR was originally developed by a collaboration of KPMG Charitable Trust with the Institute of Education and Government between 2005 and 2008. In 2008, the then-Government committed to a national roll-out of the ECaR. Due to funding cuts, the scale of the programme has been cut back.

Scope of evaluation

The research questions addressed in the reports are as follows:

- Implementation: (a) What are the strengths and weaknesses of the delivery model? (b) Has fidelity ECaR standards been consistently achieved? (c) What are the challenges to quality and sustainability?
- Impact: (a) What is the impact of ECaR on standards of literacy for eligible pupils compared to similar pupils who did not receive ECaR? (b) Are any subgroup differences observable? (c) What is the impact on whole school attainment? (d) What is the impact on wider outcomes?

Value for money: (a) What is the value for money of the ECaR programme? (b) How could the delivery model be made more cost effective?*Overall methodology*

- Implementation surveys of Local Authorities and schools.
- Qualitative case studies and interviews.
- Observation of Reading Recovery sessions.
- Impact analysis of overall initiative (ECaR) using administrative data.
- Impact analysis of Reading recovery impact study.
- Value for money analysis. A measure of cost-effectiveness is calculated based on the costs of the ECaR per pupil and the estimate of the impact. The long-term benefits of ECaR are outlined focusing on earnings , health and crime.

Impact evaluation

ECaR: Difference-in-differences (DiD) techniques to measure the impact of ECaR, exploiting the fact that ECaR policy was rolled out in stages.

Reading Recovery:
Analysis based on matching pupils in treatment schools to pupils in non-treatment schools.

Policy details

Information provided on number of Local Authorities and schools involved in Reading Recovery each year between 2005/05 to 2009/10. This increased from 205 schools (31 Local Authorities) in 2005/06 to 1,656 schools (128 Local Authorities in 2009/10). It isn't clear why these particular schools participated in the programme.

Data Administrative data from the National Pupil Database matched to pupils and schools known to be involved in ECaR and Reading Recovery respectively.

For the Reading Recovery study, questionnaires were developed for class teachers (treatment and control schools). These questionnaires covered type of literacy support received by students; an assessment of reading for each of the students involved in the study.

Costs

Costs are estimated (£3,100 per participant in the first year; £2,600 in subsequent years). The first-year cost includes the initial set-up costs whereas the cost for subsequent years does not.

Outcome variables

Reading and Writing attainment at the end of Key Stage 1 (age 7). In the second year of its operation, the ECaR improved school level reading at Key Stage 1 by between 2 and 6 percentage points. It also had an impact on school level writing. Reading Recovery had an impact of 26 percentage points on pupils reaching level 1 or above in their reading as assessed by class teachers.

Control group

ECaR impact study: The control group of schools are those that received the ECaR treatment at time subsequent to the analysis of outcomes. The treatment schools get the ECaR treatment for the first time between 2006/07 and 2008/09. The control group of schools get the ECaR treatment for the first time in 2009/10. The analysis is conducted both at school and at pupil level. In the pupil-level analysis, the sample of pupils used is below the 10^{th} (or 25^{th}) percentile of the distribution of scores of the Foundation Stage Profile (where scores are given at age 5).

Reading Recovery impact study: A stratified random sample of 153 schools participating in the ECaR programme was drawn. Comparison schools were constructed using data from the National Pupil Database and OfSTED inspections. One to one nearest neighbour propensity score matching used to match each ECaR schools to the single best comparison school. Second best matches were also found. Schools were then recruited by telephone interviewers. Schools in the treatment and comparison group schools were given guidance on how to select particular pupils for the analysis.

Methodology details

ECaR: difference-in-differences

Reading Recovery: kernel matching – matching each participant to several members of the comparison school pupil group. More weight is given to non-ECaR pupils with the most similar characteristics to the ECaR pupil.

Internal Validity:

The study was very carefully implemented.

ECaR analysis: Difference-in-differences analysis involves making a 'common trends' assumption. This is evaluated by comparing the pre-policy trends in the KS1 outcome viable for schools that implemented the ECaR in a particular year relative to a control group.

Reading Recovery analysis: Background characteristics of treatment pupils and comparison pupils are compared before and after matching. Matching greatly improves comparability.

Inference

The results are well explained.

External validity

Much care was taken to ensure that participants in the Reading Recovery evaluation were representative of those taking part in Reading Recovery more generally. However, it is not clear how representative ECaR schools are of schools generally. *Cost effectiveness*

A value for money analysis is conducted. The cost per additional child reaching the expected level at KS1 is estimated.

The lifetime benefits of the ECaR are predicted via three routes: greater earnings, better health and lower crime. Estimates are given under a 'no depreciation' scenario and a 'full depreciation' scenario. A break-even point is worked out: the impact of the programme must be sustained beyond age 11 for the policy to break-even.

Overall assessment (including suggestions for improvements – internal, external, metaphysical; useful to policy makers)?

This evaluation is extremely well done and the approach adopted would rate 4 on the Maryland scale.

If a RCT had been possible, it would have been preferable to the matching analysis (particularly for Reading Recovery). This is because we cannot be certain that the control group of pupils would have been selected for the programme, had it been introduced into these schools. This concern also applies to the evaluation of the broader programme since schools selected into the ECaR programme. However, in this case, we can see that pre-programme trends in the outcome variable were not evident between the treatment and control group.

A limitation is that the outcome variables are based on teacher assessment (Key Stage 1). It would have been preferable to have outcome variables that are based on external assessment, and at a more refined scale than Key Stage 1 outcomes. However, this is not possible when using administrative data alone.

The longer-term effects of the programme are of great interest. It is likely to be possible to do such an analysis in future years.

International comparators

Several international studies on reading recovery or similar programs designed to help struggling readers using RCTs or matching designs.

See Slavin R, Lake C, Davis S, and Madden N (2011), Effective Programs for Struggling Readers: A Best-evidence Synthesis. *Educational Research Review* 6(1), 1–26.

http://www.sciencedirect.com/science/article/pii/S1747938X10000400

Documents examined

Evaluation of Every Child a Reader (ECaR)
Emily Tanner, Ashley Brown, Naomi Day, Mehul Kotecha, Natalie Low, Gareth Morrell, Ola Turczuk, Victoria Brown, Aleks Collingwood (National Centre for Social Research)
Haroon Chowdry, Ellen Greaves (Institute for Fiscal Studies)
Colin Harrison, Gill Johnson (University of Nottingham)
Susan Purdon (Bryson Purdon Social Research)
Research Report DFE-RR114. May 2011
https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR114.pdf
Evaluation of Every Child a Reader (ECaR): Technical report

https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR114A.pdf

Key Stage 2 career-related learning pathfinder evaluation

Policy objectives

The policy context was a concern about young people's access to good quality information, advice and guidance (IAG). In 2007 The Children's Plan 14-19 Expert Group recommended that IAG should be embedded at a younger age. The Children's Plan: Building Brighter Futures (DCSF, 2007) committed the then DCSF to fund a project which would explore the impact of early career-related learning at Key Stage 2 (focused mainly on Year 6).

The Key Stage 2 career-related learning Pathfinder was a pilot programme with the following main aims: to increase pupils' awareness of career/work opportunities; increase their understanding of the link between education, qualifications and work opportunities; reduce gender specific career/role stereotypes; and engage parents/carers in the process.

Scope of evaluation

- To evaluate the extent to which the Pathfinder pilot (in 7 Local Authorities) achieved its original objectives.
- To test the hypothesis that introducing career-related learning at Key Stage 2 (in disadvantaged areas) increases and widens pupils' education and career aspirations.

Overall methodology

- Scoping study to examine implementation plans and activities: document reviews and telephone interviews with key personnel from seven Local Authorities implementing the Pathfinder pilot.
- Quantitative data collection and analysis. Involves comparing 38 Pathfinder schools to 120 (matched) comparison schools. Three surveys of the same pupils were conducted between 2009 and 2010.
- Case studies. A case study school selected in each of the seven Local Authorities – visited on 2 occasions. In total, about 60 interviews with staff and pupils were conducted on each occasion.

Impact evaluation

- Telephone interviews with key participants in Local Authorities and consultants appointed by the then DCSF. These were conducted at the beginning of the Pathfinder pilot (August-October 2009) and at the end (July-August 2010). The purpose was to find out the perception of the interviewees on the impact of the Pathfinder pilot and its sustainability.
- The quantitative analysis is described as quasi-experimental. It compares pupil responses in Pathfinder schools to those in similar non-Pathfinder schools (comparison schools). The surveys were conducted before during and after the activities had been delivered. In addition to pupil surveys, a school questionnaire was also completed by headteachers in which they were asked about career-related learning activities (completed on 2 occasions).
- The case studies involved two visits to each school, at the beginning and end of the Pilot. The first visit was to find out what career-related approaches were already used by the school, their reasons for involvement in the Pathfinder and what they proposed to do (teacher interviews) and to obtain a picture of pupils' aspirations (pupil interviews). The follow-up interviews considers how the Pathfinder was implemented and how pupils' aspirations have changed.

Policy details

LAs invited to submit proposals to deliver Pathfinder pilots across a number of primary schools within their local area. Seven LAs were selected. They are geographically spread across England, but similar in having densely-populated urban areas with high levels of social and economic deprivation. LAs invited primary schools to participate because the challenges of their social environment were considered relevant to the aims of the programme.

LAs and individual schools allowed to develop their own approach to careers-related learning but must do the following: identify their pupils' specific needs for career-related learning; audit the existing curriculum to see where this learning is already supported; design, plan and deliver a programme of careers-related learning based on the learner needs analysis and curriculum audit.

Data

The pupil survey for the quantitative analysis includes questions about what pupils' are good at, attitudes to learning, self-confidence, attitudes to school, different jobs (aimed at assessing stereotypical attitudes), the extent to which the school is good at helping to find out about different jobs, about secondary school, about university etc., helpfulness of different people (e.g. teachers, parents) on finding out about jobs; and future choices about future education and jobs.

Costs

A grant of up to £60,000 made available to each Local Authority.

Outcome variables

The quantitative analysis used items in the pupil survey to create composite measures of pupil outcomes. These themed composites were then further tested in factor and reliability analyses to check that the items correlated well with each other. The outcomes in the quantitative analysis are labelled as the following: stereotypical thinking; effectiveness of career-related learning; perceptions of parents/carers' aspirations; attitude to learning; confidence in ability to work effectively; perceived capability regarding types of career (using SOC categories for 5 different categories of job); aspirations regarding particular types of career.

Control group

The control group consists of 120 schools that were not selected for the Treatment in the 7 Local Authorities chosen for the Pathfinder. The selection of the treatment and comparison schools is not explained. Pathfinder and comparison schools are compared along a number of dimensions at baseline. A Figure is provided in the Appendix (A4.1) on sample representativenes. This includes a statement on whether treatment versus comparison schools are representative on various dimensions. However, no statistics are provided in the table so the reader is unable to gauge how comparable these groups are on observable characteristics.

Methodology details

Multi-level modelling (MLM), which takes account of hierarchical nature of the data (for example, that pupils are grouped within schools and schools are grouped within LAs). The regression analysis compares the outcomes of pupils in the treatment and control group after

accounting for a range of background variables. Also, analysis of variance (ANOVAs) which look at differences between Pathfinder and comparison schools – from the school-level questionnaire. Background variables are not taken into account in the ANOVA analysis.

Internal validity

The issue of selection into the Pilot is not addressed. The analysis assumes that comparison schools can be selected in the same Local Authorities without regard for the fact that these schools were not chosen for the Pilot by the Local Authorities. It is found that at baseline pupils in Pathfinder schools rated the effectiveness of school's career-related learning more positively than pupils in comparison schools. This indicates positive selection into the Pilot. This is not discussed in the report.

Although the analysis does look at changes over time, it does not make explicit use of this in a difference-in-differences context.

Inference

In the text, basic results are explained and there are some graphs showing changes over time in the outcome variables. The actual estimates are provided in an Appendix. However, this only gives coefficients and effect sizes. It only reports results that are statistically significant. There are so many interaction terms included in the analysis (e.g. Pathfinder status with baseline characteristics; with sweep of the survey) that it is difficult to interpret the reported coefficients – especially since variables are only reported if they are statistically significant. There are no tables that just show the difference between pathfinder and comparison group schools with baseline characteristics but no interaction terms. In the reporting of results, emphasis is given to what is found to be statistically significant but not to variables where no differential is found between the treatment and comparison groups (although attention is drawn to the fact that the Pilots did not increase the involvement of parents and carers).

External validity

This is addressed in the report only insofar as it states that pathfinder and comparison schools are representative of other schools in England in terms of school type and the percentage of pupils eligible to receive free school meals but were not representative in terms of achievement and the proportion of pupils who speak English as an additional language (EAL).The report states 'overall this suggests that the findings are relatively generalisable to similar schools in areas of deprivation, but may not entirely reflect the situation of all these schools'.

Cost effectiveness

Under value for money and sustainability, the conclusion to the report states the following: 'overall, for comparatively low costs, the case study school interviewees considered that the Pathfinder had successfully delivered on its stated aims and objectives'. The overview of the

quantitative evaluation states that the MLM 'revealed two significant overall correlations with two composite outcomes, namely generally Pathfinder pupils showed a greater decrease in stereotypical thinking and a greater improvement in their perceptions of the effectiveness of career-related learning in their school over the evaluation than did comparison school pupils'. The report also states that the Pilot helped close the gap between more vulnerable pupils and their peers.

Overall assessment

The approach adopted in this report would rate 2 on the Maryland scale because although it uses a comparison group, it does not explain how this group was selected or demonstrate that this group is comparable. Given the research design it is unlikely that the comparison group is appropriate, so it would be difficult to improve the Maryland scale rating through a better write-up (although this would be desirable, regardless).

The evaluation is very general in its stated aims and similarly general in how findings are reported. The Pilot is found to have increased pupils' awareness of career/work opportunities; understanding of the links between education, qualifications and work opportunities; and reducing gender specific career/role stereotypes. From the various methods used in this evaluation, some detail is given on how the Pathfinder achieved particular aims. However, apart from the failure of the Pilot to engage parents and carers, the report is fairly uncritical of the Pilot. Results are reported if they support the Pilot but outcomes which did not show any change between treatment and comparison schools are not discussed.

There are weaknesses in the quantitative evaluation for the following reasons: (a) there is no consideration of the selection problem; (b) the conceptual underpinning of the model is weak – particularly in how interaction terms are included; (c) only reporting variables that are statistically significant; (d) not making full use that the data afforded for comparing treatment and comparison schools over time in a difference-in-differences framework. The authors refer to the analysis as 'quasi-experimental' simply because they have an analysis that uses a treatment and comparison group. This is not how the phrase 'quasi-experimental' is typically used in the academic literature.

This evaluation suggests that along some dimensions, the Pathfinder might well have been effective and that participants respond positively to it. However, an alternative to an evaluation of this kind would have been to ask the Schools Inspectorate (OfSTED) to assess the extent to which schools implemented their careers' related learning plans according to the original proposals put forward by the Local Authority. This could have been done on the normal inspection cycle. The implementation could have been reasonably monitored since requirements were: to identify pupils' specific needs for career-related learning; audit the existing curriculum to see where this learning is already supported; design, plan and deliver a programme of careers-related learning based on the learner needs analysis and curriculum audit.

International comparators

There is a review of career guidance evaluation in the International Handbook of Career Guidance

http://www.springerlink.com/content/l2x02k0275t31356/

Documents examined

Key Stage 2 career-related learning pathfinder evaluation
Pauline Wade, Caroline Bergeron, Karen White, David Teeman, David Sims and Palak Mehta
Research Report DFE-RR116. May 2011
https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR116.pdf

Evaluation of National Citizen Service Pilots: Interim Evaluation

Policy objectives

The National Citizen's Service (NCS) is one of the Coalition Government's flagship initiatives for building a bigger, strong society. The programme aims to be rite of passage of all 16 year olds and help to promote a more cohesive, responsible and active society. The NCS involves both residential and at-home components and voluntary local action schemes.

Scope of evaluation

- To inform the future development of the NCS programme through assessment of the design and delivery of the pilot scheme.
- To assess the impact of the NCS on young people's attitudes and behaviours with regard to: social mixing, leadership, communication, community involvement and trust, confidence and transition to adulthood.
- Gather information on the views of parents of young people and the wider general public as regards NCS.
- Estimate the value for money of the NCS programme.

Overall methodology

- Inform future development of NCS: A process evaluation involving 12 case studies conducted at the NCS team level; in-depth interviews with staff and volunteers, workshops and video diaries; online focus groups; use of monitoring information data collected by providers.
- Assess impact of NCS on young people's attitudes and behaviours: Impact survey involving: baseline and two follow-up surveys of NCS participants; baseline and two follow-up surveys of matched control group from the National Pupil Database.

- Gather views of parents and general public: monitoring and analysis of print and social media content referring to NCS.
- Economic Analysis: cost benefits analysis of impacts that can be monetised; cost-effectiveness analysis of other impacts; benchmarking of NCS value for money against other programmes.

Impact evaluation

- Control group from the National Pupil Database (NPD) was surveyed using a similar baseline survey to NCS participants. A subset of this group was used as the matched comparison group to the NCS participants. The matching is done based on key socio-demographic characteristics and on attitudes to pro-social behaviour.
- The baseline survey is conducted by paper questionnaire. The follow-up is conducted by web and telephone.
- The impacts of the programme are evaluated under a wide range of variables that fall under the following headings: communication, teamwork and leadership; facilitating transition to adulthood; improving social mixing; and encouraging community involvement.

Policy details

In 2011, the programme was developed by independent charities, social enterprises and businesses, all of whom had to compete through an open tendering process to run the programme. The 2011 pilot was open to all young people around the age of 16 (who would typically have just completed year 11 or equivalent), although extended up to the age of 25 for those with learning difficulties or disabilities. In 2011, the NCS was provided by 12 organisations that made over 10,000 places available to 16 year olds in different locations across England. A total of 29 organisations have been commissioned to provide up to 30,000 places in 2012, with the aim being to raise the number of places up to 90,000 by 2014.

Data

The surveys collect information from young people on a wide range of issues - demographics; attitudes, behaviours and aspirations; and potential outcomes. Although much is reported in the text, there is no table that gives information on the main variables collected.

Costs

Costs of the NCS are provided in total and per participant. In 2011, the NCS pilots cost the government £14.2 million to deliver (with an additional £3 million raised by providers). The unit cost per commissioned place was £1,303 to government and £1,533 in total.

Outcome variables

The outcome variables in the impact analysis can be classified as follows:

a) Communication, teamwork and leadership:
   - Confidence about working with other people in a team.
   - Confidence about meeting new people, interacting with others etc.
   - Attitudes with being a leader of a team
b) Transition to adulthood
   - Personal qualities such as self-esteem.
   - Life skills – confidence in managing money and time management.
   - Progression into education, employment and training – asked about attitudes to education and plans for the future.
   - Reduction in challenging and anti-social behaviour

In each case, the quantitative findings are supplemented with insights from the qualitative research to discuss impacts under these categories.

Control group

It is stated that the control group is a sub-sample of students in the National Pupil Database. Using a baseline survey, NCS participants and non-participants have been matched based on key socio-demographic characteristics and on their attitudes towards pro-social behaviour. There is very little technical detail in this whole report (and there does not appear to be an accompanying technical report). There is one table (on measures of confidence) that shows the treatment and control group to be similar at baseline in this respect. It also shows the treatment and control group to be similar in size (about 1,500 in each) both at baseline and follow-up. However, this is the only table containing numbers from the survey in the entire interim report. No specific details are given on how the matching was done. The selection issue is completely ignored (i.e. the control group had the option of selecting into the NCS – as it was offered to all 16 year olds – but chose not to).

Methodology details

The impact evaluation matches the treatment group to a control group and then looks at the mean at baseline and follow-up. From these statistics, a difference-in-difference estimate is reported. There is only one table showing results under measures of confidence. Other findings are discussed in the text.

Internal validity

Even for an interim report (and not a technical report), there is a surprising lack of detail. From the information given, it is difficult to comment on internal validity. A major problem is obviously the fact that the treatment group selected into the NCS whereas the control group did not (although all 16 year olds had this option). This issue is not raised in the report

Inference

One table of results is reported. The effect sizes are reported with no standard errors. The other results are reported in the text. Often what happened in the treatment group and control group is reported. A comment is made about whether estimates were statistically significant.

External validity

There is a useful section of the report that describes the characteristics of all NCS participants relative to the group of participants surveyed from the NPD (before matching). The report says that this control group is weighted so that it represents the population of young people as a whole. However, no further is provided on exactly how this was done.

Cost effectiveness

The impact analysis reports a range of statistically significant positive impacts in relation to communication, teamwork and leadership; transition to adulthood; social mixing – although the overall pattern of change in this area was mixed; a small number of significant positive impacts in relation to community involvement although the overall pattern of change in this area was mixed.

The economic benefits of the programme are described as follows: benefits resulting from the time spent volunteering by the participants as part of their programme; future benefits resulting from increased teamwork, communication and leadership; and future benefits resulting from greater take up of economic opportunities.

The report states that the benefits to society as a whole are estimated to be up to £28 million. This is made up of: £618,000 in time donated by volunteers; £10.2 million in increased earnings by NCS participants because of increased confidence in teamwork, communication and leadership; and up to £17.1 million increase in earnings for NCS participants because of greater take up of education opportunities. Estimates are made of corresponding tax revenue.

The report states that as the pilot programme costs the government nearly £14.2 million, the societal benefits are between two to one times the cost of the programme.

The only information about how the monetary benefits were computed is as follows: ‘the monetary benefits are based on the best estimate available from the evaluation impact and secondary literature’.

The costs of the programme are compared to volunteer programmes like AmeriCorps, National Guard Challenge and Teen Outreach. They are shown to be in the same ballpark. The benefits are compared by ticking a box on the types of benefits provided on the NCS and these and other programmes.

Overall assessment

There are many aspects to this evaluation and the data collected should be useful to policy makers. However, the interim report is superficial with regard to the impact study and the economic evaluation. It gives much too little detail and there is no reference to a technical report (nor is there one on their website). As it stands, the report would rate 2 on the Maryland scale, because it does not demonstrate that the comparison group is appropriate. In principle, if more work was done to improve to demonstrate comparability, this report could rate 3-4 on the Maryland scale.

Some of the findings reported are interesting. However, since the selection problem is not dealt with, comparisons between treatment and control groups cannot be taken to reflect the causal impact of the NCS (except under strong assumptions).

The monetary estimates of benefits are very hard to believe. Strong assumptions have clearly been made to translate soft skills and future intentions to young peoples' behaviour and success in the labour market.

A careful descriptive analysis of the programme would be better than what has been produced. The matching analysis might be interesting if carefully described. Full transparency is necessary with regard to methodology and reporting of results. The cost-benefit analysis should be done making much more conservative assumptions and these should be set out clearly.

International comparators

There is a World Bank report that reviews impact assessments of youth voluntary service programs. It also outlines best practice.

http://siteresources.worldbank.org/INTCY/Resources/395766118789951541 4/ReportYouthServiceMeeting.pdf

Documents examined

Evaluation of National Citizen Service Pilots: Interim Report
NatCen Social Research, The Office for Public Management, and New Philanthropy Capital
Date: May 2012
Prepared for: The Cabinet Office

http://www.natcen.ac.uk/media/898405/ncs-evaluation-interim-report.pdf

Social and emotional aspects of learning (SEAL) programme in secondary schools: national evaluation

Policy objectives

SEAL is ‘a comprehensive, whole-school approach to promoting the social and emotional skills that underpin effective learning, positive behaviour, regular attendance, staff effectiveness and the emotional health and wellbeing of all who learn and work in schools’. At the time of this report, it was implemented in around 90% of primary schools and 70% of secondary schools.

SEAL is designed to promote the development and application to learning of social and emotional skills that have been classified under the following five domains: self-awareness, self-regulation, motivation, empathy, social skills.

SEAL is envisaged as a loose enabling framework for school improvement rather than a structured package that is applied to schools. Schools are encouraged to explore different approaches to implementation rather than follow a single model – so SEAL is what individual schools make of it.

Scope of evaluation

- To assess the impact of secondary SEAL on a variety of outcomes for pupils, staff and schools.
- To examine how schools implemented SEAL with particular reference to the adoption of a whole-school approach.

Overall methodology

- Pupil-level surveys in treatment and comparison group schools to assess the impact of SEAL.
  26 SEAL schools and 23 comparison schools.
- Qualitative study primarily to provide insights into the implementation process (and also used to discuss impact). A subset of the SEAL schools from the quantitative evaluation (10 schools) used for this purpose. Case study schools visited 5 times. Data collection comprised observations of lessons and other contexts; interviews and/or focus groups with members of the school community; and analysis of school documents.

Impact evaluation

- 22 SEAL schools drawn from the secondary schools that were initially selected by their Local Authority for the initial roll-out (about 300 in total) of this national programme. This started in October 2007 and the schools had declared that they intended to implement the programme from this point forward.
- 19 comparison schools drawn mostly from the same Local Authorities as the 22 SEAL schools. They had chosen not to implement the SEAL programme and this was checked each year of the study (2008-2010).
- Pupil surveys were conducted in all schools (Year 7) and the treatment and comparison group compared using Multi-Level Modelling (i.e. comparing treatment and control schools after controlling for a range of characteristics; method accounts for hierarchical

nature of the data: schools clustered within Local Authorities; pupils clustered within schools).

Policy details

Initial roll-out of secondary SEAL took place in 2007/08 (about 300 schools targeted initially). The programme became national very quickly – implemented in 70% of secondary schools by the time of the final report in 2010.

Data

Pupils surveyed three times: at the beginning of 2008, 2009 and 2010 respectively. A range of data collected on social and emotional skills and general mental health as well as administrative data held at school level (and in the National Pupil Database) about pupils.

Costs

In the qualitative evaluation, lack of time and resources comes up as a barrier to implementation. It is stated that most schools received little or no financial resources to aid implementation, which meant that simple needs such as being able to buy relevant media and prepare lessons resources was problematic for some.

Outcome variables

- Pupil self-report version of the Emotional Literacy Assessment Instrument.
- General mental health difficulties, pro-social behaviour and behaviour problems as measured by the pupil self-report version of the Strengths and Difficulties Questionnaires.

Control group

Comparison group of schools selected mainly from the same Local Authorities as the subset of treatment schools that agreed to take part. After the treatment group had been established, comparison schools which shared similar observable characteristics (in administrative data) were approached. The treatment group was drawn from the 300 secondary schools that were targeted by their LAs for the initial roll-out. However, SEAL is a national programme and comparison schools had selected not to implement this programme.

Methodology details

Regression analysis where pupils in SEAL schools are compared to pupils in the control group (using Multi-level modelling). The results are discussed clearly in the text with the full tables presented in the Appendix.

Internal validity

No discussion of the selection bias which is inherent to the design (i.e. control schools had selected not to implement the treatment). The researchers call this approach quasi-experimental – apparently because they use a control group.

Inference

In general, results are clearly explained in the text, with the full results and models given in the Appendix.

External validity

As discussed in the report, the flexible nature of the programme means that each school can take a very different approach to the implementation of SEAL. As the authors state, it is very difficult to make generalisations about the success or failure of SEAL overall. However, the report has a table which comparison schools participating in this study with the national average. This is interpreted by the authors to indicate that schools are 'broadly similar'. However the statistics indicate that they are lower performing and contain a higher percentage of students eligible to receive free school meals.

Cost effectiveness

The analysis of outcomes suggests little change in treatment schools relative to control schools. The mean values for each outcome measure are reported for the baseline survey and the final survey. Controlling for other variables in the regression analysis makes little difference.

There is no discussion of costs. In the qualitative evaluation, it is apparent that schools received little or no funding for implementing the programme.

Overall assessment

The quantitative analysis has the merit of being transparent – the results are clearly reported and discussed. However, the 'quasi-experiment' has been misconceived – it is not possible to make inferences from a comparison group that self-selected not to undertake the treatment. As such, the report would rate 2 on the Maryland scale (because it fails to demonstrate that the comparison group is valid). The research design (with self-selection in to treatment) might make it hard to demonstrate comparability and improve the Maryland scale rating.

The qualitative analysis is very thorough and enables a critical discussion of this policy, putting the findings in the context of related literature. However, in the literature review, there were five other studies of the SEAL programme discussed. This included three studies relating to primary SEAL and two relating to an earlier pilot of the SEAL programme in secondary schools. Many of the findings of this study support earlier studies (including one by OfSTED – the Schools Inspectorate). It is not clear why there had to be another specially commissioned evaluation of SEAL when OfSTED would have been able to monitor

implementation during the usual schools inspection cycle. Furthermore, since schools are encouraged to develop their own programme, the results of a small survey were never going to be generalisable to other schools.

International comparators

The PENN Resiliency Programme is another programme that tries to increase child wellbeing. This has been evaluated in the US and in England. DfE have also commissioned this work. It lends itself to an impact evaluation because it is a pilot rather than a national programme. The treatment and control groups were decided as part of the evaluation process.

https://www.education.gov.uk/publications//eOrderingDownload/DCSF-RR094.pdf

Documents examined

Social and emotional aspects of learning (SEAL) programme in secondary schools: national evaluation. Neil Humphrey, Ann Lendrum, Michael Wigelsworth. DFE-RR049. October 2010
https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR049.pdf

The impact of Sure Start Local Programmes on five year olds and their families

Policy objectives

The ultimate goal of Sure Start Local Programmes (SSLPs) was to enhance the life chances for young children growing up in disadvantaged neighbourhoods. The aim was to bring together early education, childcare, health services and family support to promote the physical, intellectual and social development of babies and children. They were targeted to specific disadvantaged areas and all children living in the targeted area and their parents were eligible to receive services.

Scope of evaluation

There are various different components to the National Evaluation of Sure Start Team: core team; impact module; implementation module; cost-effectiveness module; local context analysis module; support to local programmes on local evaluations module; data analysis team.

This evaluation relates to the 'Impact of Sure Start Local Programmes on Five Year Olds and their Families' and the corresponding report 'National Evaluation of Sure Start local programmes: an economic perspective'.

- To measure the impact of SSLPs on children and their families when the children are five years old.

- The economic study estimates costs and benefits of SSLPs and discusses potential future benefits.

Overall methodology

- The impact analysis uses matched treatment and control areas. There are four stages of analysis
  - Whether there are across-the-board effects of SLLPs on child and family functioning when children were 5 years of age or in terms of change over time in the case of outcomes measured at both 3 and 5 years of age.
  - Whether effects detected by comparing the treatment and control group samples might have under or over-estimated impacts (by considering outcomes in areas outside the 'common support')
  - Whether effects of SSLPs vary across demographically defined sub-populations.
  - An analysis of the possible impact of attrition.
- The economic evaluation looks at what SSLPs cost; potential economic benefits that might arise from measured outcomes; sources of potential long-term economic benefits; predicting long-term economic benefits; conclusions about the short-term and long-term impact.

Impact evaluation

- Intention-to-treat design – measure impact of being in an eligible area for SSLP as compared to being in a matched comparison area.
- The main analysis investigates the effect of SSLPs on child development and family functioning. The outcome variables are grouped under: Child Behaviour and Social Development; Child Physical Health; Child Educational Development; Maternal Wellbeing and Parent and Family Functioning.
- Data analysed using multilevel models, which take into account the hierarchial structure of the data, with children and families nested within communities. Linear models are used for continuous measures and logistic models for binary outcomes. The results compare children and families in areas eligible for SSLP compared to those in the MCS control group.

Policy details

The first 524 Sure Start local programmes (SSLPs) were established between 1999 and 2003. The services (childcare, family support) were made available to all children under the age of five living within designated areas. Initially SLLPs did not have a prescribed curriculum or set of services. Instead each SLLP had extensive local autonomy over how it fulfilled its mission to improve and create services as needed, without specifying how services were to be changed. From 2005-2006, fundamental changes were made in SLLPs as they came under the controls of LAs and operated as children's centres. This modified the service-delivery process in that the guidelines were more specific about the services to be offered.

Nonetheless there is still substantial variation among LAs and areas within LAs in that way that the new model is implemented.

Data

Extensive surveys conducted. For the treatment group, information was collected by a specially trained fieldworker (home visit lasting about 90 minutes) when children were 9 months and again at 3 years of age and 5 years of age. MCS data collection done in a similar way. There is some information which is provided by the teacher – the Foundation Stage Profile which covers six areas of learning.

Costs

Collect information on the cost of SSLPs from four sources: regular financial information provided by Sure Start local programmes to the Sure Start Unit from 1999-2000 to 2004-05; information from the implementation surveys of Sure Start local programmes; information from implementation case studies; information about children's centre expenditure from the National Audit Office report on Sure Start Children's Centres.

Sure Start local programmes cost an average of £4,860 (including capital costs) per eligible child living in the area at 2009-10 prices over the four years that children and their families were eligible to receive services. There was substantial variation around this total – the highest spending SSLP spent more than £12,000 per eligible child and the lowest spending sent less than £2,000.

The report is unable to measure the overall take up rates for services on a consistent basis, and thus calculate expenditure per child who actually use SSLP services.

Outcome variables

About 20 outcome variables are included in the main table. The come under the categories Child Behaviour and Social Development (e.g. emotional dysregulation, positive social behaviour, self-regularly), Child Physical Health (e.g. BMI), Child Educational Development (as measured in the Foundation Stage Profile across all schools), Maternal Wellbeing (e.g. mother's satisfaction with life, self-rated depression), Parent and Family Functioning (e.g. health discipline in home; chaos in home; home learning environment).

Control group

The control group is from a matched comparison group of children in the Millennium Cohort Study (MCS) in areas not covered by SSLPs. The treatment group is a randomly selected subset of children and families previously studied (at 9 months and 3 years) for an earlier evaluation of Sure Start. The data pertains to 5 year old children in each case. However, the fieldwork for the MCS was done two year prior to the fieldwork for the treatment group (ending March 2007 and June 2009 respectively).

Methodology details

Propensity Score Matching using children and families in the SSLP follow-up sample, as compared to the MCS sample. The surveys are all of children and families that have been interviewed on several occasions up to when the children were aged 5 years. However, the MCS sample is interviewed 2 years earlier.

The MCS areas were carefully selected (and the methodology is explained in detail). Propensity Score Matching was conducted at the area level – matching on 85 indices of deprivation and other socio-demographic variables obtained from administrative sources. The data were divided into 5 strata where stratum 1 was least likely to be chosen as a SSLP area (relative advantage) whereas stratum 5 were most likely to be chosen (most disadvantaged). Because of 'common support' issues the treatment-control differences can only be considered for those in stratum 2-4.

Internal validity

The analysis is carefully conducted. However, as discussed in the report, there is potential confounding of year effects with the effect of the treatment because the control group is surveyed two years before the treatment group. There is also selective attrition within the treatment and control group (who were surveyed by two different teams of researchers). The report examines the potential of attrition in the treatment group to obscure the results (the 5 year old sample is a subsample of those interviewed when the child was age 3). The report rejects this possibility because on some measures the former sample is more disadvantaged whereas on others it is less disadvantaged.

Inference

Most details of results are presented. However, standard errors of estimates are not provided (the 95% confidence intervals are shown and an indicator of whether estimates are statistically significant).

External validity

There is some investigation about whether the estimated effects might be generalizable to areas outside the 'common support' of the treatment and control groups. This is done by comparing the outcomes of children/families across 'stratums' within the treatment group (where the most disadvantaged area – and most likely to be in the treatment – had no similar area which could serve as a control). The report states that the government decision to double the number of SSLPs meant that few communities without an SSLP remained.

Cost effectiveness

The main positive effects for children relate to health - those in the treatment group had lower BMI and better physical health. There are four positive and two negative effects on outcomes

relating to maternal wellbeing and family functioning. There are no differences between the treatment and control group on seven measures of cognitive and social development from the Foundation Stage Profile. There is also a reduction of the proportion of children living in families where no parent was in paid work in the treatment group relative to the control group.

The economic study suggests that the impact relating to worklessness could translate into a short-term economic impact whereas longer-term potential economic benefits come via lower rates of conduct problems and higher educational attainment. The reduced conduct problems are related to the lower probability of committing crime in the future. The report suggests that the impact indicators at age five (less home chaos, less harsh discipline and a better home learning environment) are all associated with lower rates of worklessness as adults and lower rates of reoffending. However, they interpret effect sizes to be small and do not attempt to translate these effects into potential longer-term outcomes.

The report monetises the effects arising from the fact that parents in eligible areas move into work more quickly. The report estimates this to be between £279 and £557 per eligible child. This is compared to costs of around £1,300 per year for each eligible child (£4,860 over the period from birth to the age of four). The report emphasises that benefits may not become apparent for 10-15 years. However, since the early evidence suggests no impact on the Foundation Stage Profile, it is unclear where these large gains in educational attainment are expected to come from (and why they are not apparent in the FSP).

Overall assessment

The analysis is very carefully done and well reported. The limitations are understood and explained. Overall, the approach adopted would rate 4 on the Maryland scale.

However, one fairly obvious thing to do would have been to analyse differences of a treatment and comparison group within the MCS instead of using a different data source for the treatment group (measured 2 years later). This might have been rejected because of sample size considerations. However, it was never discussed in the report. There is ongoing work at the Institute of Education (not part of the evaluation team) using the MCS for this purpose.

It is also not clear why the control group used in the early evaluation of SSLP was not followed up (this might have been on account of cost considerations). However, it would have been useful to obtain information from different cohorts and then use this data in the same analysis. The report does make comparisons with the early evaluation (which found more mixed results of SSLP), but it does not appear to be possible to use different cohorts of treatment and control in the same analysis.

SLLP did take time to roll out – and it is unfortunate that the timing across areas could not have been randomised such that a more robust evaluation would have been possible. A particular issue is that it was not possible to measure effects for those living in the most

deprived areas because there was no suitable comparison group. This is unfortunate from a policy perspective because effects on the most disadvantaged communities might have been of greatest interest.

International comparators

Evaluation of Head Start. This was commissioned by the US Department of Health and Human Services. This has a RCT design.

http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/impact_study/executive_summary_final.pdf

Documents examined

The impact of Sure Start Local Programmes on five year olds and their families. The National Evaluation of Sure Start (NESS) Team, Institute for the Study of Children, Families and Social Issues, Birkbeck University of London. DFE-RR067. November 2010.

https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR067.pdf

National evaluation of Sure Start local programmes: An economic perspective. National Evaluation of Sure Start Team led by Pam Meadows. DFE- RR073. July 2011.

https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR073.pdf

Randomised controlled trial of the 'Teens and Toddlers' programme

Policy objectives

To assess the impact of the 'Teens and Toddlers' (T&T) youth development and teenage pregnancy prevention programme. This trial forms part of a wider evaluation that included a stage of formative qualitative work and a process evaluation.

The aims of the T&T programme were to decrease teenage pregnancy by raising the aspirations and educational attainment of 13-17 year old teenagers at most risk of leaving education early, social exclusion and becoming pregnant. The programme was implemented through secondary schools and involved three-hour weekly sessions in a nursery setting for 18 to 20 weeks. Each participant supports a child, takes part in classroom-based group work, keeps a journal of their experience and has access to a trained counsellor.

Scope of evaluation

RCT intervention with the following steps:

- At-risk young women identified by their teachers using guidance provided by T&T
- Individual girls randomly allocated to a treatment or control group
- Data for participants collected by questionnaire at three points in time.
- Two cohorts of girls participated in the trial – one starting in Sept 2009 and one in January/February 2010. In total 449 teenagers participated (228 in the treatment and 221 in the control).

Overall methodology

- Collect baseline data of teenagers taking part in the RCT and analyse characteristics of those assigned to the treatment and control groups.
- Analysis of the views of participants about their experience of the programme (collected via an additional questionnaire).
- Impact analysis from the RCT.
- Discussion of methodological limitations; interpretation of findings.

Impact evaluation

- Within school RCT (22 schools) and 449 participants.
- Four primary outcomes of interest and 14 secondary outcomes. The primary outcomes are: did not use any contraception the last time they had sex (and had sex within the last 3 months); has had more than one episode of not using contraception in the last three months; expects teenage parenthood, youth development score (made up of selected items from the 'youth at risk' version of the Life Effectiveness Questionnaire). The secondary outcomes consist of a range of outcomes relating to contraception, self-esteem, difficulty in discussing sex, pregnancy and school absence. An additional primary outcome is added to tables (not discussed in the Executive Summary) – a youth development score which is made up of
- The average age of participants was 13.5 years: 44% receiving free school meals; 32% living in workless households; 2% had been pregnant; 13% had experienced sex.
- Impact analysis is on 'intention to treat' – all teenagers originally assigned to treatment and control groups were analysed, regardless of how many sessions of the T&T programme they attended in total.
- Follow-up conducted immediately after the programme finished.

Policy details

T&T is a youth development and teenage pregnancy prevention programme that aims to decrease teenage pregnancy by raising the aspirations and educational attainment of 13-17 year old teenagers at most risk of leaving education early, social exclusion and becoming pregnant. It focuses on geographical areas with high rates of teenage pregnancy. There is discussion about the context of this programme and that it is a central project of a charity called 'Children: Our Ultimate Investment'. However, there is no discussion of the extent to which this programme is applied in the UK. The schools were selected by people in the T&T

programme and primarily consisted of schools with which they had established working relationships.

Data

Surveys of girls in treatment and control groups. Changes between cohorts 1 and 2 in the baseline survey – from computer assisted personal interviews to paper questionnaires (as the latter was thought potentially better for disclosing sensitive information). There were big differences in the baseline survey between cohorts 1 and 2 on reporting of sexual behaviour. The change in survey format is one potential explanation; another is that 10 schools took part in both cohorts; 12 schools only took part in the second cohort. There is relatively little attrition in the surveys (and the attrition does not vary between the treatment and control group). However, a significant proportion (in both the treatment and control groups) attended less than half of the programme (about 25% overall).

Costs

Not discussed

Outcome variables

Four primary outcomes of interest and 14 secondary outcomes. The primary outcomes are: did not use any contraception the last time they had sex (and had sex within the last 3 months); has had more than one episode of not using contraception in the last three months; expects teenage parenthood, youth development score (made up of selected items from the 'youth at risk' version of the Life Effectiveness Questionnaire).

Control group

'At risk' girls grouped into matched pairs within schools. Then one member randomly allocated to the T&S programme and the other to the control group. Reserve pairs were chosen in the event of drop-out within the first 8 weeks of the intervention. The matched pairs were chosen based on age and sexual experience. There was a risk of contamination of the control group because of communication between those in the treatment and control groups. This is discussed by the researchers but they did not have sufficient resources to increase the sample size and then conduct a cluster RCT (which would have avoided this problem).

Methodology details

RCT with a baseline survey and 2 follow-up surveys: immediately post-intervention and one year after the intervention.

Internal validity

The design of the experiment ensures internal validity. However, had the researchers needed to use too many 'reserves' (i.e. pairs of students to be included in the programme if others dropped out within the first 8 weeks), this might have been misleading about the effects of the programme. However, they only needed to do this in 10% of cases.

One risk to the internal validity of this experiment is contamination between the treatment and control group due to girls (within the same school) talking about it to each other.

Inference

In general results are well reported. However, it seems strange to adjust the control group for baseline differences between the two groups in the main table of results (though a more detailed table is given in the appendix). In a RCT, it would have been more appropriate to show means for the control group, intervention group and differences. Then an additional table could have shown some regressions to see the treatment-control difference after controls are added.

External validity

No discussion of this.

Cost effectiveness

Costs of the programme are not discussed.

The intervention was not found to have had any effect on the primary outcomes of interest. However, there was evidence of a positive impact on 3 of the 14 secondary outcomes: self-esteem; knowledge of sexual health; less likely to report difficulty in discussing the pill with a doctor.

Overall assessment

This is a well implemented evaluation with a strong methodological design. The overall design would rate 5 on the Maryland scale, although the trial is on a relatively small scale and there are problems outlined below which potentially reduce its rating. There are some limitations of the study which are discussed by the authors. However, they attribute the programme's lack of success to features of the programme (e.g. insufficient sexual health education) rather than to the conduct of the evaluation.

Potential problems with the evaluation are as follows: Firstly, there appears to be a lower than expected prevalence of (acknowledged) risky behaviour in the sample used in the study. If one compares the control group in the study with the expected rate in the control group (i.e. the basis on which detectable effect sizes were estimated in advance), the actual control group displays less evidence of risky behaviour. This would have made it difficult to find effects in some of the outcome variables (particularly failure to use contraception).

Furthermore, a significant number of the treatment group (25%) attended under half of the sessions – and these girls were the most likely to be from disadvantaged backgrounds and engage in risky behaviours. Secondly, there was a risk of contamination between the treatment and control group. A clustered RCT would have been better (though a more expensive project). It would also have made it possible to design an intervention that would test the effect of peer-to-peer interaction in this context. Thirdly, the study is fairly short-term.

International comparators

There have been a number of RCTs aimed at reducing unintended pregnancies among adolescents. This are reviewed in the following article:

DiCenso, A., Guyatt, G., Willan, A., and Griffith, L. (2002), 'Interventions to reduce unintended pregnancies among adolescents: systematic review of randomised controlled trials'. *British Medical Journal*, 324, pp.1426-34. This does not find positive evidence for any of the primary prevention strategies tried and tested.

Documents examined

Research report:
Randomised controlled trial of the 'Teens and Toddlers' programme.
Ruth Maisey, Svetlana Speight, Peter Keogh and Ivonne Wollny NatCen Social Research
Chris Bonell, Annik Sorhaindo and Kaye Wellings London School of Hygiene and Tropical Medicine
Susan Purdon Bryson Purdon Social Research. DFE-RR211. May 2012.https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR211.pdf

Appendix: Evaluations in the area of Labour Market policy

This appendix provides details of the evaluations considered in the area of labour market policy. The structure of the template was agreed following discussions with the National Audit Office. In completing the templates, for reasons of both feasibility and presentation, we have made use of source material from the original evaluations without any attempt to provide detailed attribution (e.g. through the use of quotes, or the provision of page numbers).

Employment Retention and Advancement Demonstration

Policy objectives

Overall, ERA aimed to intervene decisively in the 'low-pay, no-pay' cycle, whereby low-skilled and disadvantaged workers move frequently between low-paid work and out-of-work benefits, and to turn them, instead, into regular full-time workers. Evaluation was built in as an integral part of the programme, which featured large scale random assignment, and was overseen by DWP.

ERA targeted three groups with different views on, and preparation for, work and advancement:

- 'The NDLP group': Unemployed lone parents receiving Income Support1 and volunteering for the New Deal for Lone Parents (NDLP) welfare-to-work programme;
- 'The WTC group': Lone parents working part time and receiving Working Tax Credit (WTC), which supplements the wages of low-paid workers;
- 'The ND25+ group': Long-term unemployed people aged 25 or older receiving Jobseeker's Allowance2 and who were required to participate in the New Deal 25 Plus (ND25+) welfare-to-work programme.

Scope of evaluation

The evaluation is divided into three main research strands:

- A process study: relies on qualitative and quantitative data, intended to provide insight into possible reasons for the programme's impacts or lack of impacts.
- An impact study: This study uses administrative records data and customer surveys to compare the service receipt, employment, earnings, benefits receipt, and other outcomes for ERA participants with those of the control group members.
- A cost-benefit study: examines the net economic gains or losses (or net present value) generated by ERA by comparing the costs of the programme with the financial benefits it induces.

The 'final evidence' report discussed here covers all these strands, although the process evaluation is not reviewed. There are numerous other evaluations and working papers which

report on specific target groups, earlier stages of the intervention, aspects of the process, and costs.

Overall methodology

Randomised control trial. Qualifying members of the three target groups were invited to volunteer for a fixed number of ERA openings that would be allocated on a randomised basis. After completing an informed consent process, half of the volunteers (there were over 16000 volunteers) were assigned randomly to the ERA programme group, and the rest to a control group. Those in the control group could continue to receive whatever services they were normally entitled to receive from Jobcentre Plus or could obtain elsewhere in the community. ERA's success was determined by comparing the outcomes of the programme group, such as average earnings, with the outcomes of the control group.

Impact evaluation

Estimates impacts on various labour market outcomes for each of the target groups: work and earnings, employment dynamics and job characteristics, training, individuals' steps to advancement, benefit receipt, and wellbeing. Reports on differences by region, differences across Job Centre Plus offices, by sub-groups (education, ethnicity) and on long-term impacts.

Policy details

ERA was implemented in 6 regions: London, East Midlands, North East England, North West England, Scotland, Wales. Launched in 2003 in selected Jobcentre Plus offices, which administer Government cash benefits and employment services, the programme was envisioned as a 'next step' in British welfare-to-work policies. Participants in ERA had access to a distinctive set of 'post-employment' job coaching and financial incentives, which were added to the job placement services that unemployed people could normally receive through Jobcentre Plus. Once employed, ERA participants could receive at least two years of advice and assistance from an employment adviser to help them continue working and advance in work. Those who consistently worked full time could receive substantial cash rewards, called 'retention bonuses'. Participants could also receive help with tuition costs and cash rewards for completing training courses while employed.

Data

- Impact analysis uses data from DWP administrative data covering all ERA programme and control group members. Employment and earnings administrative records data were provided to DWP by HMRC and maintained in DWP's Work and Pensions Longitudinal Study (WPLS).
- Customer survey administered by phone or in person to a sample of programme and control group members, at 12 months, 24 months and 60 months after randomisation. Response rates varied from 93% (WTC group, 12 month) down to 62% (NDLP 60 month).
- In-depth qualitative interviews with staff and programme group members from 2004 through spring 2009. Weekly diaries of participant contact from advisers.
- Staffing and salary data, plus DWP administrative data on incentive payments to participants for cost analysis.

Costs

Estimates of the cost of the ERA advisers and clerical personnel who provided the services estimated using the advisers' time diaries and staffing form data, and imputed overheads (using district specific weightings). Similar steps were followed in estimating the staff cost of serving the ERA control group, including the cost of Personal Advisers for controls. Costs

estimated for 33 months of services and financial incentives to an average participant in each of the three target groups.

Outcome variables

- Employment: ever employed; average number of months, employed months 24,36,48,60; number of employment/non-employment spells; months to first employment; duration of first employment; time to first employment). Hours worked, from survey data;
- Earnings: by year, and 2005-2009;
- Job characteristics: permanent; shift work; time off; responsibilities; supervision; opportunities for promotion; views about work; trade union;
- Training: participated in training (by year 1-2, 3-5); obtained qualification; participation in training by full-time/part-time work status; number of training courses; hours in training;
- Advancement: e.g. tried to increase hours of work, get pay rise, go to career office; sign up with recruitment agency;
- Benefits: Job Seekers Allowance, Income support, Incapacity benefit, Housing Benefit, Working Tax Credit, Child Tax Credit, e.g. amounts, number of months received, ever received;
- Wellbeing: subjective wellbeing, health and financial outlook, child's performance and wellbeing at school;

Control group

The design involved explicit randomisation. Half of the individuals in the three target groups who volunteered for the programme were assigned at random – regardless of their background characteristics – to a programme group that was enrolled in ERA or to a control group that was not enrolled in ERA. The randomised treatment/control group design is central to the evaluation method.

Methodology details

Random assignment of ERA volunteers to treatment and control groups. Results presented are group means, differences and p-values. In some cases (not always clear) estimates are regression adjusted for control/treatment group characteristics.

Internal validity

Internal validity relies on effectiveness of randomisation in balancing treatment and control group characteristics. Final report lacks any comparison of the pre-treatment characteristics or outcomes of the treatment and control groups, except in one or two specific instances where the pre-treatment comparison is made graphically, and for the 60-month survey sample. There are potential attrition problems in the survey data (but not the administrative data).

Inference

P-values for differences in means reported. No details on method, e.g. whether clustered at JC+ offices or not. The evaluation reports many different tests, of which only a few in the NDLP and WTC groups are statistically significant. Potential issues from multiple comparisons not discussed. P-values not reported for comparison considered 'non-experimental' i.e. where comparison is made between a non-random subset of the randomised samples (e.g. those working)

External validity

Administrative data covers PAYE employees and not self-employed. Attrition in survey data in later years may make it non-representative, although various checks reported in Appendix A. Representativeness of regions and offices selected for study, and of those who volunteered to participate in the study is not discussed in the final report, with no comparison with corresponding populations. There was evidently a problem in the way the study was administered which meant that, as well as the 'refusers' who declined to participate, a high proportion of those who should have been eligible were not offered the programme. Therefore 23-30% of the target population are not represented (Sianesi, 2010). Sianesi (2010) and Chowdry and Sianesi (2011) present working papers on the external validity of the ERA study. The conclusion is that although the participants are not always representative of the target populations, this either does not affect the findings, or leads to the impacts being underestimated (ND25+ and NDLP groups show +ve employment and earnings benefits). These issues do not appear to be discussed in detail in the final report. Timing issues have been covered in detail with consideration of short versus long term impacts. Potential displacement (e.g. if employment or earnings gains came at expense of others with less labour market coaching) versus additionality issues are not considered and would be hard to address.

Cost effectiveness

Evaluation presents a detailed cost benefit analysis, using standard discount rates. Inevitably, net benefits are sensitive to the assumptions used. Most of the significant impacts are on the ND25+ group (baseline +£2500 per person over 10 years) due to earnings impacts. Negative net benefits for the NDLP (-£107) and WTC (-£1600) groups, although finds positive effects for A-level qualified NDLP. Costs are 'up and running' costs and do not include costs of setting up or evaluating ERA.

Overall assessment

This is a large scale evaluation with numerous sub-reports and working papers, carried out by multiple organisations (IFS, PSI, NIESR, DWP, ONS, MDRC) over a period of 10 years. The programme design and part of the analysis was carried out by the US organisation MDRC which designs similar policy experiments in the US. The final report is large and complex (300 pages), and comes against a background of 16 other reports and working papers related

to the programme. The design of the evaluation (randomised control trial) is ideal in principle, although there were evidently problems in implementation which left the study participants non-representative of the target populations. This evaluation would count as a Level 5 on the Maryland scale, and, setting aside its potential costs is a good model for programme evaluation. A large part of its strength, other than in showing the effectiveness of the intervention on long term unemployed, is in demonstrating the application of large scale randomised control trial methods to social policy in the UK. The evaluation itself was presumably very costly, which is a potential drawback of an evaluation of this type, although no information on these costs is provided in the final report. There are some weaknesses in the final report in that it lacks pre-treatment group comparisons, technical details of the methods are not always clear, and the volume of the final report and the body of literature surrounding the ERA makes it quite inaccessible. Questions of displacement, while not easy to address, deserve some attention. The summary of the final report and its conclusions is however clear and succinct.

International comparators

The programme and its evaluation were based on a similar programme in the US, the US Employment, Retention and Advancement project, also carried out by MDRC.

Documents examined

Hendra, Richard, James A. Riccio, Richard Dorsett, David H. Greenberg, Genevieve Knight, Joan Phillips, Philip K. Robins, Sandra Vegeris, and Johanna Walter, with Aaron Hill, Kathryn Ray, and Jared Smith (2011), Breaking the low-pay, no-pay cycle: Final evidence from the UK Employment Retention and Advancement (ERA) demonstration, DWP Research Report http://research.dwp.gov.uk/asd/asd5/rports2011-2012/rrep765.pdf

Chowdry Haroon. and Barbara Sianesi (2011) Non-participation in the Employment Retention and Advancement Study: Implications for the experimental fourth-year impact estimates, DWP Working Paper No. 96 http://research.dwp.gov.uk/asd/asd5/WP96.pdf

Sianesi, Barbara (2010), Non-participation in the Employment Retention & Advancement study: Implications for the experimental first-year impact estimates, Department for Work and Pensions Working Paper No 77. http://research.dwp.gov.uk/asd/asd5/WP77.pdf

European Social Fund

Policy objectives

The European Social Fund (ESF) was set up to improve employment opportunities in the European Union and so help raise standards of living. Its aim is to help people fulfil their potential by giving them better skills and better job prospects.

Scope of evaluation

Evaluation focuses on participants who entered the programme between June 2008 and April 2009 and estimates the programme impacts on two broad DWP customer groups: participants in receipt of Jobseeker's Allowance and participants in receipt of Incapacity Benefit or Employment Support Allowance.

Overall methodology

Comparison of voluntary participant and non-participant groups, using individuals matched by propensity score on characteristics available in administrative data.

Impact evaluation

Quantitative work provides impact analysis on various benefit and employment outcomes.

Policy details

The study is focused on the Department for Work and Pensions (DWP) ESF funded employment provision part of the programme, which was contracted by DWP during 2008-11, and delivered by private, public and third sector providers at an expected cost of £265 million. A key feature of ESF funding is that it must be used to purchase additional provision in order to extend coverage, address gaps and complement domestic funding. The provision itself is varied and flexible, including activities such as job search guidance, basic skills training, case worker support and advice on tackling specific barriers to work. Participation by individuals in the programme is voluntary. There are 3 general categories:

- Tailored (a flexible, personalised approach) - 51 contracts, cost £190m;
- Targeted (contracts in which provision is specified to particular needs – for example helping participants with English language barriers or participants with a disability) - 19 contracts, cost £70m;
- Intermediate Labour Market (high unit cost contracts for providing subsidised temporary employment with the aim of providing a bridge back to the labour market) – four contracts, cost £5m;

Data

Administrative data from DWP and HMRC

Costs
Programme costs stated (see above)

Outcome variables

Receipt of benefits and employment.

Estimated separately of JSA and IB or employment support allowance claimants.

Control group

Comparison groups of non-participants in receipt of JSA and IB/ESA are drawn from the population of individuals who could have entered the programme during the same time period as participants in the sample. Groups of non-participants are selected who most closely resemble ESF participants with regard to demographic characteristics, benefit and employment history and prior participation on DWP programmes.

Methodology details

Estimates average effect of treatment on treated using propensity score matching (kernel based 1-1 matching using psmatch2). Individuals selected on to programme by Jobcentre plus or through own choice, so selection can depend on adviser and individual characteristics. Matching based on age, gender, ethnic group, disability, qualification, marital status and lone parent status, occupation choice, IMD, labour market history indicators

Internal validity

Limitations of matching methods discussed in relation to selection on unobservables. Common support and balancing tested, and sensitivity tests discussed.

Inference

Graphical presentation with confidence intervals, but methods for estimation unclear.

External validity

Estimates based on administrative data likely to be representative, but the results are for average effect of the treatment on the treated so not necessarily representative in the presence of heterogeneous treatment effects. The period corresponds to the recession, so generalizability to different economic conditions is unknown (this point is discussed). Diverse range of interventions makes it hard to generalise to other future interventions.

Cost effectiveness

The evaluation does not provide and explicit cost effectiveness or cost benefit calculation, but the impact estimates could be used to derive these (with additional information).

Overall assessment

A carefully done evaluation, which takes into account many of the issues typically discussed in the modern matching based evaluation literature. The design is fundamentally limited by comparison of voluntary participants and non-participants, since these unobservable selection effects are almost certainly present. The timing of the evaluation during the recession is unfortunate, as it raises questions about the generalizability of the findings. These issues are noted in the report. Level 3 on the Maryland scale in that differences between treatment and controls are likely to remain uncontrolled for, although there are elements of Level 4. The evaluation is report is clear and concise. A limitation of the evaluation is that the range of interventions provided under ESF contracts are diverse so it is not at all clear what intervention is being evaluated and what lessons can't be learnt for future interventions.

International comparators

There is a large volume of experimental evidence on similar programmes in the US on mandatory and voluntary programmes aimed at job seekers, see Appendix B of Card, David, Jochen Kluve and Andrea Weber (2010) Active Labor Market Policy Evaluation: A Meta-Analysis, Economic Journal, 120(548) F452-F477. There is a European Commission funded evaluation of the ESF across Europe, but this is based on case studies, interviews, expert opinion and aggregate descriptive statistics.

Documents examined

Ainsworth, Paul and Simon Marlow (2011) Early Impacts of the European Social Fund 2007-13, DWP In House Research IHR3 http:// research.dwp.gov.uk/asd/asd5/ih2011-2012/ihr3.pdf

Fair Cities Pilots

Policy objectives

The Fair Cities Pilots in Birmingham, Brent and Bradford made up an experimental programme, which aimed to increase the number of disadvantaged ethnic minority residents who gain steady work and new careers. The programme aimed to test the effectiveness and value-for-money of the 'demand-led' approach in tackling disadvantage in the labour market. It is an area-based active labour market programme. There were some specific targets 4424

job entries, 65% from disadvantaged wards, 70% of interviewed participants starting work, 70% employment retention at 13 weeks)

Scope of evaluation

Two phases, Phase 1 2005, Phase 2 2005 on. Phase 1 Evaluation aimed to provide process evaluation, assess the feasibility of quantitative and comparative evaluations, largely through qualitative methods. Phase 2 aimed to assess short and long term impacts on beneficiaries, employers and local employment, training and community 'infrastructure'.

Overall methodology

This is an evaluation based on qualitative methods, with some descriptive numerical information on numbers of job entries, characteristics of participants. The methods involved:

- Case study qualitative research with Pilots and the Central Secretariat;
- Qualitative research with stakeholder and community organisations;
- Case study qualitative research with employers;
- Qualitative research with providers;

Impact evaluation

No explicit evaluation, because no counterfactual. The evaluation is based primarily on the views of participants and stakeholders.

Policy details

Involved nebulous 'pipelines' for matching people with specific vacancies. Efforts to identify or create 'employer-led' or 'demand-led' vacancies, through the involvement of local business leaders in 'Local Board' meetings (with JC+ and other representatives) and promotion to programme participants. Promoted training by existing providers to prepare 'disadvantaged jobseekers for specific vacancies and their associated personal and skill requirements'. Beneficiary eligibility was not restricted to members of ethnic minority groups but targeted at wards within the cities which high ethnic minority and inactive. The report notes that any job gains were likely to be displacement, with 'substitution in favour of their target beneficiaries at the expense of other jobseekers'. Participation in the 'Pilots' was voluntary for both employers and potential beneficiaries, participants in mainstream New Deal programmes were ineligible, and JSA claimants restricted by work limit rules.

Data

Administrative data on programme participation and outcomes. Qualitative information from fieldwork. The methods for the qualitative work are not described in the evaluation reports.

Costs

Provides summaries of overall expenditures in the three years up to 2008 (£9 million).

Outcome variables

Reports figures on job entries. Narrative on design and operation of the system, characteristics of jobs offered, characteristics of participants and those obtaining jobs, reasons for employer engagement, roles of training providers.

Control group

None used.

Methodology details

The evaluation provides descriptive evidence on programme participants, number of jobs started and other outcomes, plus qualitative evidence based on interviews.

Internal validity

The evaluation is not explicitly making many causal claims, and is mainly a description of the activities of stakeholders, their views and characteristics. Misleadingly, figures are presented for numbers of job creations and costs per job, although none of the evidence indicates that these jobs would have been created or that participants would not have found employment without the intervention.

Inference

No explicit statistical hypothesis tests.

External validity

The findings are not easily generalizable to any other context or programme.

Cost effectiveness

The evaluation provides an indication of cost effectiveness in terms of costs per job created (around £9000). These costs are compared with the costs of other ALMP programmes from other sources, with which it compares very unfavourably. There is some discussion of benefits based on reported one year benefits from other forces, although this seems to compare one off costs with annual benefits from employment, and is unclear.

Overall assessment

This evaluation is not a programme evaluation in the usual sense, although it provides a basic analysis of cost effectiveness and costs per job ‘created’. The overall evaluation is conveyed

through the presentation of researchers' opinion based on observation and interviews. These arguments seem plausible, and the evaluation is critical of the programme costs and impacts, but the ways in which these opinions were reached are not always set out transparently. In general it fails to provide specific evidence on what the programme delivered relative to any baseline or business-as-usual scenario. The qualitative discussion of the delivery of the programme and the attitudes of stakeholders is informative, and potentially useful in delivering lessons to policy makers engaged in the design of similar programmes. There is, however, inadequate information on how the interviews were conducted, and how interviewees were selected, so the generalizability is questionable. The evaluation of the quantitative impacts of the programme on the target group is weak. It acknowledges various limitations: e.g. that job creation was potentially displacement and that the volunteer participants were not representative of the target group, but no reasons are given for not employing alternative evaluation approaches using comparators from the target group. As an evaluation, this rates Level 1 on the Maryland scale.

International comparators

Any area-based programme evaluation. No directly related international studies known.

Documents examined

Atkinson, John, Sara Dewson, Harriet Fern, Rosie Page, Rachel Pillai and Nii Djan Tackey, Evaluation of the Fair Cities Pilots 2007 DWP Research Report 495 http://research.dwp.gov.uk/asd/asd5/rports2007-2008/rrep495.pdf

Gateway to Work

Policy objectives

Objectives of GtW were to:

- Increase the numbers of people moving into jobs in the early stages of Gateway;
- Reduce the level of Gateway overstayers;
- Improve the participants' motivation and ability to participate in their Intensive Activity Period (IAP);
- Improve participants' readiness for employment.

The GtW pilots took place in four areas: London (across every District), Manchester, Swansea and Dundee. These formally ended in March 2006.

Scope of evaluation

Qualitative work by GHK and quantitative work on administrative data by DWP, aims to identify and explore the impact of the pilots; identify best practice, in terms of which

elements of GtW have been most effective in moving clients into employment. One chapter on impact evaluation, one on cost benefit; most of report is qualitative.

Overall methodology

Qualitative fieldwork; four case studies. Quantitative evaluation is difference-in-difference estimation comparing pilot offices and areas with selected comparator offices. Methods are not explained in detail e.g. no information on how pilots were chosen.

Impact evaluation

Quantitative work provides impact analysis on various benefit and employment outcomes.

Policy details

GtW is a two week, full-time training programme which is mandatory for ND25+ clients that have been claiming Jobseeker's Allowance (JSA) four weeks after joining Gateway. The course provides soft skills training in areas such as communication, team building and problem solving as well as CV writing, interview techniques and support with applying for jobs.

Data

No details provided on qualitative data work, or on administrative data.
Costs
Information provided in cost effectiveness analysis (see below).

Outcome variables

Exits and length of time on New Deal and destinations.

Control group

Offices with similar labour market characteristics (i.e. in the same Jobcentre Plus cluster) to the pilot offices which had no or very low (less than ten per cent) GtW referrals.

Methodology details

Nothing to add to methodology outline above.

Internal validity

Unclear how comparison offices selected. Treatment and control groups reported as balanced on gender, age and length of claim, but not on ethnicity and disabled. Evaluation states that "Overall, the comparison offices were felt to be similar enough to the pilot offices to allow

robust results to be obtained". Descriptive statistics on balancing, but no statistical tests. No results on sensitivity to assumptions. Groups do not appear balance on pre-treatment employment or benefit status.

Inference

Some confidence intervals reported, but many of the results are without this information. No information provided on clustering assumptions.

External validity

The results are for treatment on the treated i.e. potentially only valid for office of the type piloted.

Cost effectiveness

Provides a basic cost effectiveness analysis – both fiscal (saving in benefits minus cost of programme) and 'economic'. Latter assumes that £1 in taxes reduces economic output by 25% (basis for this unclear). Cost savings appear to be due to deterring claimants from staying on benefits and moving to 'intense activity period', with no employment gains. These savings do not outweigh the costs.

Overall assessment

Has the potential for a good evaluation design using comparison of treatment and control areas in a pilot program, although effectiveness of the piloting for evaluation unclear, since no details provided on how pilots chosen (presumably not random). This is not a very detailed evaluation, and seems quite hurried. The quantitative work manually selects comparator areas and the basic difference in difference method is appropriate, but there is insufficient detail on how the comparisons made. The treatment and control areas are not effectively balanced and no steps are taken to address this problem. Potentially Level 2-3 on the Maryland scale, although basic in implementation. Comes to a clear (and plausible) conclusion that this programme was not value for money, although the evidence presented is not very robust.

International comparators

There is a large volume of experimental evidence on similar progammes in the US on mandatory and voluntary programmes aimed at lone parents see Appendix B of Card, David, Jochen Kluve and Andrea Weber (2010) Active Labor Market Policy Evaluation: A Meta-Analysis, Economic Journal, 120(548) F452-F477.

Documents examined

Page, James, Dr Eleanor Breen and Jayne Middlemas Gateway to Work New Deal 25 Plus pilots evaluation, DWP research report 366 http://research.dwp.gov.uk/asd/asd5/rports2005-2006/rrep366.pdf

Job Centre Plus

Policy objectives

The original business case for Jobcentre Plus set out a list of 12 deliverables against which to justify the investment. Crucially, the business case was based on the assumption that Jobcentre Plus would increase effective labour supply leading to an improvement in the functioning of the labour market, with consequent economic benefits and public expenditure savings. Specifically, it assumed that once up and running, Jobcentre Plus would move more than 140,000 people in the hardest-to-help groups into work every year. However, net additionality – the net change in total employment – was assumed to be only 28,000 (20 per cent), once account was taken of substitution and displacement effects. Additional job outcomes generated in this manner were expected to result in Annually Managed Expenditure (AME) savings of £620 million per annum, just under 60 per cent of the total annual savings to the Exchequer associated with the investment in Jobcentre Plus.

Scope of evaluation

The main objective of the analysis in this report is to assess the labour market impacts of the introduction of Jobcentre Plus. Specific questions addressed are:

- What impact has the introduction of Jobcentre Plus had on the numbers of people moving off benefit and into work?
- What impact has the introduction of Jobcentre Plus had on the employment rate overall and the employment rate of different sub-groups?
- What impact has the introduction of Jobcentre Plus had on the wider economy, including output and the public finances?

The evaluation is of JC+ relative to previous arrangements.

Overall methodology

Estimation of impacts from gradual roll out of programme across geographical areas between 2001 and 2005/6. Multiple event difference-in-difference. Simulation of policy effects from NIESR macro model. The difference-in-difference approach is applied both to aggregate labour market flow data at Job Centre District Level and to individual level data (for those coming into contact with JC+). Treatment in the JCD flow estimates is measured by job centre "intensity" i.e. "Percentage of the stock of JSA claims that are registered in offices where Jobcentre Plus is live" (?)(p.65). Treatment in the individual data is making a new claim for JSA.

Impact evaluation

Estimates impact on exit rate from benefits to work (over 3 month period), and inflows and outflows into specific benefits (JSA, IS, disability). Investigates heterogeneity by client group (over 50, disabled etc.). Also looks at impacts on wages. Predicts impacts on unemployment and employment using these flow estimates. Simulates macro effects using NIESR macro model.

Policy details

Before 2002 public employment services were delivered through the Employment Service. The job-brokering activities and active labour market policies provided through the Employment Service were directed primarily at people claiming unemployment benefits (Jobseeker's Allowance (JSA)). Separately, a range of social security benefits were provided through the Benefits Agency, including Income Support (IS) and Incapacity Benefit (IB) ('inactive' benefits).

Jobcentre Plus was first introduced in 56 Pathfinder sites in 17 clusters across the UK in October 2001. The second stage of implementation, known as Day Two and covering 24 districts, began in October 2002 and was mostly completed in March 2003. The remainder of the national roll out was scheduled in three successive waves between 2003/04 and 2005/06.

The roll-out of Jobcentre Plus, which involved a £1.9 billion spend, represented a major overhaul of the infrastructure used to deliver public employment and benefit services. The main change was to bring the Employment Service and Benefits Agency under one roof, providing an integrated service for all people of working age seeking social security benefits and involving a significant rationalisation of estates.

Data

Secondary data from mainly administrative sources. Aggregated benefit claimant outflows from NOMIS (derived from administrative data). Department for Work and Pensions' (DWP) National Benefits Database (NBD) to develop different policy indicators to capture the 'treatment' effects of Jobcentre Plus and for individual level analysis. Labour Force Survey used for wage estimates.

Costs

Evaluation of effects on costs (i.e. transfer payments) provided by NIESR macro model. Costs of programme roll-out includes, although sources unclear.

Outcome variables

Exits from various types of benefit to employment. Exits from JSA to other benefits. Employment stock. Unemployment stock. Wages. Public finances.

Control group

Individual analysis: treatment group is individuals making a new claim in JC+ areas, control group is individuals making a claim in selected comparator areas; post policy period defined by start of claim 6-9 months after introduction of JC+ ("Day two phase only"); pre-policy period defined by start of claim 15-18 months before JC+ role out. Various comparator areas investigated including all non "Day Two" areas, and subset based on pre-policy characteristics. Propensity score matching for individuals in individual claimant analysis.

There is no explicit separate control group in the aggregate analysis, although low or zero JC+ penetration districts implicitly provide the controls for those with high penetration.

Methodology details

The individual level analysis is a detailed econometric study, which demonstrates consideration of a wide range of potential threats to identification.

The JCD flow estimates are based on an error correction model, to allow for temporal dynamics. It is not completely clear how the standard (Nickell 1981) problems in dynamic panel data models have been addressed. Aggregate (across JCD) exit rates are used instead of time dummies or time specific trends to control for macro shocks.

The individual level analysis is the most finely tuned in terms of methods and testing, but it appears to be the results from the aggregate level analysis that feed into the main evaluation conclusions.

Internal validity

No robustness tests presented for the ECM model of claimant flows. Alternative specifications are presented for the individual claimants analysis, based around simple difference-in-difference versus a 'random growth' model to allow for differential trends, versus difference-in-difference with propensity score matching. No presentation of sensitivity to inclusion or exclusion of controls in either estimation procedure. There are potential issues with timing of intervention because high-performing offices were chosen for Pathfinder stage. Numerous tests for balancing and differential trends carried out.

All approaches suffered from differences in pre-policy trends between treatment and comparator groups. Differences corrected by 'random growth' difference-in-difference models (i.e. adjusted for differences in trends). In a second approach, individuals picked from non JC+ areas using propensity score matching. Extensive tests of matching of samples in individual data. No specific choice of comparator groups or tests of balancing presented in aggregate flow models.

Jobcentre Plus was first introduced in 56 Pathfinder sites in 17 clusters across the UK in October 2001. The second stage of implementation, known as Day Two and covering 24

districts, began in October 2002 and was mostly completed in March 2003, with the remainder of this stage of the roll-out being completed over the year that followed.22 The Pathfinder sites represent six per cent of the population claiming benefits and Day Two districts represent 23 per cent. The remainder of the national roll out was scheduled in three successive waves between 2003/04 and 2005/06.

This timing of the introduction of JC+ provides the basis for the quasi-experimental design used to generate estimates of its impacts. Timing issues appear to have been considered carefully and in great detail in the analysis, with estimates of both long run and short run impacts.

Inference

Confidence and intervals and standard errors presented throughout, although no details of methods (e.g. clustering levels) is provided.

External validity

Administrative data and Labour Force Survey data provides good external validity, subject to sample selection rules. Displacement from other areas not addressed. Potential for JC+ recipients to exit benefits to employment at expense of non JC+ groups is not specifically addressed.

Cost effectiveness

Evaluation includes an explicit statement of costs and about the effect of JC+ on public finances. These estimates are derived by simulation from the NIESR macro-economic model, using the labour market impacts estimated from the other sections of the report. The gains come from the increased flow off benefits and a 0.1% improvement in GDP. The report claims JC+ will have had net positive effect on public finances by 2015 of about +5.5 billion.

Overall assessment

The individual level estimation method appears very robust, but finds relatively weak or non-existent effects. The aggregate panel data analysis is good in principle, although the analysis is complicated by using time-series based ECM model, in order to derive short run and long run impacts. The main report conclusions appear to derive mainly from the aggregate flow models, although these are potentially much less robust than the micro data estimates and no thorough sensitivity testing or assessment of the aggregate flow models is presented. The linkages between the micro and aggregate analyses are not completely transparent.

The net benefit calculations provide useful headline figures. It is difficult to assess the credibility of these, given they on the aggregate flow models and no details given on the underlying macro model used to generate them.

A high quality evaluation overall that has made a serious attempt to estimate causal effects. There is a mixture of quality in terms of research design. The individual analysis is potentially a 4 on the Maryland scale, given that the roll out provides a quasi-experimental research design, but has a low weight in the overall evaluation findings. The aggregate analysis is closer to level 2, although would rank higher with more careful tests for balancing and pre-treatment trends. The final cost benefit calculations rely on simulation from a macro model, and so are only as good as the underlying model. The report is long, dense, but clear on detailed reading.

The report could have been improved by providing better linkage between the micro and aggregate level analyses, and justification for preferring the aggregate flow analyses in making the final conclusions. More detailed testing the aggregate flow models in terms of balancing, sensitivity to specification.

International comparators

Not aware of any directly related studies on JC+ equivalents, but Appendix B of Card, David, Jochen Kluve and Andrea Weber (2010) Active Labor Market Policy Evaluation: A Meta-Analysis, Economic Journal, 120(548) F452-F477 provide international comparisons on ALMPs generally.

Documents examined

Rebecca Riley, Helen Bewley, Simon Kirby, Ana Rincon-Aznar and Anitha George (2011) The introduction of Jobcentre Plus: An evaluation of labour market impacts, by, Department for Work and Pensions Research Report No 781
http://research.dwp.gov.uk/asd/asd5/rports2011-2012/rrep781.pdf

Job Outcome Targets

Policy objectives

Scheme involves changing the way in which the targets for Jobcentre plus offices (and hence presumably DWP) are defined, rather than any specific interventions. Policy objectives are not defined in the evaluation, although implication is that the objective was to reduce costs.

Scope of evaluation

Qualitative and quantitative evaluations. This report deals with quantitative work carried out in-house by DWP.

Overall methodology

Difference-in-difference, based on comparison of pilot districts with comparator districts, or individuals in pilot districts and individuals in comparator districts. Pilot and comparator districts were matched by jobcentre "cluster", percent "integrated" (these terms not defined) and absence of other reforms.

Impact evaluation

Investigates the impact of the policy on off-flows from benefits, differences across client groups, and impacts on JOT staff activity.

Policy details

Not completely clear from evaluations, but appears to have involved redefining the Jobcentre targets to movements off benefits to employment (JOT), rather than reports of filled vacancies (JET). Two pilot groups were defined: which these differ in the way outcomes were recorded, option 1 using DWP/HMRC work and pensions longitudinal study, and option 2 using local "existing processes" (not explained). Pilots were implemented in Jan 2005.

Data

Administrative data from the Work and Benefits Longitudinal Survey

Costs

Discussed in relation to costs per job calculations and value for money.

Outcome variables

This is very hard to understand as none of the graphs are labelled to indicate what the outcome is and the text does not explain. The graphical analysis appears to present 'outcomes' in percentages normalised to 100 in a base year. It is possible that these are related to the moves off benefits into employment, but this is not specified.

Control group

Districts selected by matching to pilots (method unclear). Individuals in pilot districts matched by characteristics to individuals in control districts (again method unclear).

Methodology details

The evaluation is lacking detail on the matching process, although the basic method is as outlined above. The analysis is presented graphically, and appears to be looking for breaks in the trends around Jan 2005, which is a sensible design, but the methods for testing for these differences are not set out clearly.

Internal validity

Unclear on how pilot areas chosen, and on methods to match control districts and individuals. No tests presented for robustness or sensitivity to assumptions. No balancing tests for treatment and control groups.

Inference

Graphical presentation, with some parts labelled as statistically significant. Vague details on using 1% confidence intervals to compensate for "non-sampling error" variation in the data. Most of the analysis seems to rely on visual comparisons.

External validity

Administrative data, so good on representativeness of the study areas; but necessarily only representative of these areas not others.

Cost effectiveness

The evaluation presents a 'value for money' analysis, based on changes in estimated cost per job figures.

Overall assessment

This is an unusual evaluation in that is looking for evidence of any negative impacts from policy that is intended to reduce administrative costs, rather than looking for impacts from an intervention that involves spending money. The basic design is appropriate, but the evaluation is written in a way that the details would only be understood by the people who wrote it or others close to the programme in DWP. It is badly put together and appears rushed. Lacks critical evaluation of its own methods. Missing important details about the methods used, which makes detailed assessment impossible. From the report it is simply impossible to deduce what are the outcome measures. The underlying research design is potentially Level 3 on the Maryland scale but implementation is poor.

International comparators

No known comparators of this type of policy change, although other ALMP programme evaluations are relevant.

Documents examined

Frankham, John, Laura Payne, Phillip Smith, Dan Sturman and Rob Willis (2006) Evaluation of the Job Outcome Target Pilots: quantitative study Final report, DWP Research Report No 316

New Deal for Disabled People

Policy objectives

NDDP was a voluntary programme designed to help people with disabilities and health conditions move from incapacity benefits into sustainable employment.

Scope of evaluation

The overall evaluation has eight components:

- Impact analysis.
- Cost-benefit analysis.
- Documentary analysis and survey of Job Brokers.
- Survey of the eligible population.
- Qualitative research with participants, Job Broker staff and Jobcentre Plus staff.
- Survey of registrants.
- Qualitative research with employers.

- Survey of employers

Around 20% (40 pages, from 200) of the final evaluation is devoted to the impact and cost benefit analysis. Generally, the evaluation of the New Deal for Disabled People is designed to establish:

- The experiences and views of NDDP stakeholders, including Job Brokers, participants, the eligible population, employers and Jobcentre Plus staff.
- The operational effectiveness, management and best practice aspects of the Job Broker service.
- The effectiveness of the Job Broker service in helping people into sustained employment and the cost effectiveness with which this is achieved

Evaluation was directed by CRSP at Loughborough, but the impact analysis was contracted out to Abt Associates Inc and the Urban Institute in the US.

Overall methodology

Overall evaluation uses mixed methods based on field survey, interviews for qualitative content, administrative data. The impact analysis uses administrative data. The impact analysis estimates are non-experimental but based on an exact/discrete matching design.

Impact evaluation

Impact evaluation and cost benefit analysis is based on comparison of NDDP registrants with a matched control group using administrative data. Other quantitative aspects of the evaluation are descriptive, or use basic regression analysis. These parts of the analysis investigate the decision to enrol in NDRP by comparing a survey of the eligible population with a survey of NDRP registrants, and looks at characteristics of registrants and job brokers associated with movements into work and length of employment.

Policy details

NDDP is available to persons claiming one of a number of benefits related to disabilities e.g. Incapacity Benefit; Severe Disablement Allowance; Income Support with a Disability Premium. NDDP is delivered by a network of around 60 Job Broker organisations, which are a mixture of voluntary, public and private sector organisations, which help clients find work e.g. by helping clients with job search, to engage in job development, raising client confidence. The relationship of this policy to other New Deal programmes occurring simultaneously, and the implications for the evaluation, are not discussed.

Data

Impact analysis uses administrative data provided by Department for Work and Pensions (DWP) on individuals eligible for NDDP on the basis of receipt of Incapacity Benefit (IB), Income Security with Disability Premium (IS-DP), or Severe Disablement Allowance (SDA) benefits during the first three years of the programme, between July, 2001 and June, 2004. Additional data are as follows:

- Surveys of the eligible population and registrants were carried out in 3 waves. These surveys contribute to robustness tests of the impact analysis, and to descriptive and regression-based evidence on NDRP participation and registrant outcomes.
- Qualitative interviews with participants, job brokers, Job Centre Plus staff and Disability employment advisers.
- Cost data were collected from a small sample of 20 Job Brokers, which is around 20% of those initially questions (other Job Brokers unwilling or unable to provide costs). There is some censoring due to unwillingness of brokers to reveal true costs.

Costs

Job Broker cost analysis includes staff costs, overheads, and payments to other organisations by Job Brokers and is used to estimate costs per registrant. Additional costs included for Job Centre Plus administration to arrive at costs per registrant, placement and 6 month 'sustainment'.

Outcome variables

Average net impacts on more recent and longer-term claimants in each cohort were estimated for the following six outcomes for each month of the relevant follow-up period:

- Receipt of Incapacity Benefit, Income Support with disability premium or Severe Disablement Allowance (that is, incapacity-related benefits);
- Monthly amount of combined Incapacity Benefit, Income Support with disability premium and Severe Disablement Allowance;
- Receipt of Jobseeker's Allowance;
- Monthly amount of Jobseeker's Allowance;
- Rate of employment;
- Proportion of the follow-up period employed.

Control group

A control group is formed by exact matching of non-registrants to registrants based on discretised individual characteristics available in the DWP administrative data.

Methodology details

Impact analysis uses discrete cell, exact matching design, in which NDDP registrants are matched to up to 10 non registrants based on a set of (discrete) individual characteristics: registration/start month; years of prior benefit receipt (including all spells); type of IB received at registration/starting month; age at registration/starting month; sex; type of disability; DWP administrative region; (for new claimants) month in which benefit receipt began; and (for new claimants) whether there was a work-focused interview. This comparison group is then re-weighted to adjust for differing numbers of matches. Non-registrants are assigned a synthetic registration date, based on time on benefits and calendar date. Estimates are carried out for various registrant groups: an early cohort registering 2001-2002 (with a subset from July-2001 to Dec 2002 labelled maximum follow up); late cohort registering July2004-Dec2004. Further sub-group analysis carried out on the early-cohort on differences in response to treatment. Other parts of the evaluation are qualitative, present descriptive statistics, or simple regression analysis.

Internal validity

Exact-matching design relies on assumption that matching variables are sufficient to control for differences between groups (like regression, propensity score matching). Sensitivity of the design to omitted variables was tested using a field survey data (interviewed as part of the evaluation's Survey of Eligibles and Wave 1 Registrant Survey) that provided additional

information on race/ethnicity, qualifications, occupation and industry of previous employment, household composition, personal health status and functional ability. There is a tension between the survey based evidence on the decision to participate in NDRP and the claim that the matching procedure is adequate: NDRP registrants were evidently more likely to be previously looking for work in work or expecting to work in the future, showed a pre-intervention dip in activity (Ashenfleters dip), were less likely to have partners, less likely to have an employed partner. The impact and cost benefits analysis includes prediction of long run benefits from out of sample regression predictions (from 24-36 months up to 82 months, based on a quadratic). This is very bad practice, and the post-36 month impacts are potentially unreliable.

Inference

Indication is given of statistical significance for the key impacts. There is graphical analysis where statistical significance is not explicit. It is not clear whether any adjustment was made to standard errors to allow for correlation within job brokers or districts (e.g. clusters).

External validity

The impact analysis is based on administrative data, and concerns a national policy, so scores well in terms of external validity. Data on costs is less reliable, being based on small censored samples. The survey data on eligible population and registrants suffered from a high level of non-response and attrition (50-60% response for eligible population, 60-80% for registrants, depending on wave) and is weighted for non-response (based on admin data sampling frame).

Cost effectiveness

A separate report provides a detailed cost effectiveness and cost benefit analysis and takes into account range of potential costs, benefits from reductions in costs, and benefits from increased earnings. The procedure for imputing the monetary benefits from employment crude but typical of this kind of analysis (increase in average duration is estimated from % increase in employment per month and multiplied by the mean earnings of employed NDRP recipients). Benefits past 36 months from registration are predicted out of sample from a quadratic regression. Censoring of costs has impacts on the reliability of the Job Broker cost estimates (£800 to £1000 per person). The finding is, roughly, that the scheme generates no net benefit for the customer gains to the government, and (hence) societal gains (up to £3000 per year for long term claimants in £2005 prices, £600-800 for short term). The estimates assume that the jobs created were new, and not displacement.

Overall assessment

The evaluation has several components, qualitative and quantitative, which generates a fairly lengthy and complex overall evaluation. The final evaluation report draws on other reports which contain much of the detail on the methods and surveys, so it is necessary to cross

reference multiple reports to fully assess the methods used. The bibliography for the final report cites around 20 evaluations, preliminary reports, synthesis reports and component evaluations carried out by or for the DWP in relation to NDDP between 2001 and 2007.

This assessment focusses primarily on the impact and cost analysis, which is of good quality, given the available data, and has been carried out by US agencies with expertise in the field. There are notable limitations in the matching methods used. The validity of the estimates relies on the decision to register for NDDP being random, conditional on a fairly limited set of matching characteristics. The method will fail if the decision to register is based on unobserved personal characteristics which are correlated with subsequent labour market outcomes, as the field survey evidence suggests. In general matching on observables is unlikely to imply matching on unobservables when programme participation is voluntary. The method may not be a significant improvement on standard regression methods and it would be helpful if standard (OLS) regression results were presented, so this comparison could be made. The evaluation does make a serious attempt to deal with and test for selection biases, and is open about the limitations. Robustness tests based on the subsample with additional survey data indicates that there are unobserved differences between the matched groups in the main impact analysis, and controlling for these additional characteristics leads to a substantial (1/3rd) reduction in the estimates. The impact analysis is limited to benefit receipt and employment status lacks evidence on earnings and other labour market outcomes. Overall, the impact evaluation is Level 3 on the Maryland scale. The qualitative and descriptive sections of the evaluation, which account for most of the evaluation material score 1-2.

The cost evaluation and CBA is done carefully, given the data available, although has its limitations: costs are based on small sample, benefits are based on employment effects only plus cost reductions, and use out of sample (quadratic) trends to predict long run (post 36 month) employment effects.

International comparators

Wide range of active labour market programme evaluations available internationally. Related international disability-specific programmes are described in DWP in house report 90. The evaluation is comparable or better than other recent evaluations of disabled persons labour market policy internationally, e.g. Human Resources and Skills Development Canada (2010) the Evaluation of the Canada-Manitoba Labour Market Agreement for Persons with Disabilities, Human Resources and Skills Development Canada. Project Network in the US provided a randomised control trial study of ALMP for disabled people in the early 1990s (references provided in Corden 2002).

Documents examined

Orr, L., Bell, S. and Lam, K. (2007). Long-term impacts of the New Deal for Disabled People, DSS Research Report No. 432. Leeds: CDS.

Greenberg, D. and Davis, A. (2007). Evaluation of the New Deal for Disabled People: cost and cost-benefit analyses. DWP Research Report No. 431.Pires, Candice, Anne Kazimirski, Andrew Shaw, Roy Sainsbury and Angela Meah, (2006) New Deal for Disabled People Evaluation: Eligible Population Survey, Wave Three DWP Research Report No 324

Corden, Anne (2002) Employment Programmes for Disabled People: Lessons from research evaluations, DWP In House Report 90

Bruce Stafford et al (2007), New Deal for Disabled People: Third synthesis report – key findings from the evaluation Department for Work and Pensions Research Report No 430 http://research.dwp.gov.uk/asd/asd5/rports2007-2008/rrep430.pdf

New Deal for Lone Parents

Policy objectives

In 1998, the Government introduced NDLP nationally as one of a range of policy initiatives aimed at lone parents. The programme aims to 'encourage lone parents to improve their prospects and living standards by taking up and increasing paid work, and to improve their job readiness to increase their employment opportunities'.

Scope of evaluation

An initial evaluation was carried out by NatCen (Lessof, Miller et al. 2003) and a re-evaluation was commissioned to test the robustness of the previous report's conclusions:

- to evaluating the findings of the previous study, particularly the matching process, definition of participation window and non-response:
- produce impact estimates for key variables such as numbers off benefits, and numbers into jobs and report on the possible range within which the true impact of NDLP lies;
- examine longer-term outcomes of NDLP up to the end of 2003.

Overall methodology

The original NatCen report applied propensity score matching to novel large scale survey data of the eligible population and subsequent participants. The re-evaluation replicates this method, but explores sensitivity to changes in the matching design, other econometric refinements, and more general issues like equilibrium effects.

Impact evaluation

Impact evaluation was one part of the original evaluation (20 pages out of 120 pages in the original report), which also contained descriptive evidence on the reasons for participation

and non-participation and experiences of the programme. The re-evaluation refers to the impact evaluation component of the original report.

Policy details

Advice provided through advisers to:

- Encourage and motivate all lone parents to identify their skills and develop confidence;
- Provide support and guidance to clients in finding and applying for jobs;
- Improve awareness and knowledge of and provide support with benefits and tax credits;
- Help with the transition from claiming Income Support (IS) into work by providing 'better off' calculations, assisting with in-work benefit claims and liaising with employers;
- Identify and support access to education or training courses with a 'direct' work outcome to increase job readiness;
- Provide practical support in finding childcare, applying for child maintenance and liaising with the Child Support Agency;
- Offer in-work support.

The NDLP participant is not given any extra IS or other benefit than they would otherwise be eligible for, but they may be eligible for financial help with travel costs to attend job interviews, childcare costs or fees for training courses.

Data

Complex data based on survey carried out by NatCen. Postal survey in 2000 of 65000 NDLP-eligible, NDLP non-participants based on administrative records (stratified sample). Identification of subsequent NDLP participants during 2000-2001 and matched non-participants from this sample given follow up face to face interviews. Sampling based on geographical units TTWAs. Original sample complemented with more recent extracts from administrative sources (the Labour Market System and the Generalised Matching Service).
Costs

Neither the original NatCen evaluation nor the re-evaluation provides any information on costs of the programme. The re-evaluation raises the need for information on costs and CBA in the conclusion.

Outcome variables

- Exit from Income Support ;
- Rates of entry into work (both overall, and by the nature of the work) ;
- Perceived barriers to work;
- Rates of take-up of training;
- Awareness of benefits and tax credits.

Control group

Control group in original NatCen data obtained by propensity score matching of NDLP participants with non-participants.

Methodology details

NatCen evaluation compared mean outcomes of NDLP participants and matched non-participants using bespoke propensity score matching algorithm. Re-evaluation repeated the analysis using a different matching algorithm and testing the robustness along many other dimensions.

Internal validity

Detailed assessment of the internal validity of the original evaluation is one of the main goals of the re-evaluation. Their overall finding is that the internal validity of the original design was generally satisfactory, although there is little evaluation of the internal validity and robustness in the original evaluation. There were ways that the matching could have been improved – specifically on pre-treatment benefit histories - and balancing of the treatment and control samples was imperfect. The impacts were found to be sensitive to the specification of NDLP treatment duration and multiple spells, but NDLP duration is dependent on individual characteristics and the outcome of the treatment (i.e whether NDLP resulted in a move off benefit), so it is arguable whether more detailed treatment of duration offers an improvement.

Inference

The original evaluation presented means for the treatment and control groups, but in most cases no standard errors or p-values and no detailed information about methods for statistical testing, apart from the note that ‘testing of differences of this magnitude is rather academic’. Re-evaluation reports standard errors, although no detail on e.g. clustering assumptions.

External validity

Assessment of external validity is a second goal of the re-evaluation, and one of the main conclusions is that the results using administrative data (better external validity) are substantially different to (smaller than) those of NatCen using the postal survey. The selective response in the original interview sample seems to have been a problem, but the original solution of weighting for non-response made little difference. External validity of the re-evaluation should be good considering it uses administrative data. Both reports note that any estimates of this type are context dependent (refer to a specific period and country, in the context of other contemporaneous ALMP). The second report notes that there may be spillover effects on the control group (‘general equilibrium effects’, implicitly including displacement effects) although does not address them.

Cost effectiveness

Both studies provide a range of impact evaluations that could be used to generate cost effectiveness and cost benefit analyses, but neither report undertakes this analysis.

Overall assessment

The original impact evaluation involved substantial data collection and novel sampling design, and the evaluation method was adequate although was lacking in assessment of its assumptions and not very detailed. The re-evaluation provides a detailed assessment and extension of the previous impact evaluation, focussing on the econometric detail. It is not clear why this re-evaluation was commissioned, other than that the original estimates seemed unusually high, the impact evaluation was lacking in detailed robustness checks, and that there was clearly scope to extend the analysis which was a relatively small component of the original. The general conclusion of the re-evaluation is that the original evaluation estimates were indeed too large (by a factor of 2), although there is no obvious specific flaw in the original design leading to this. The lower estimates in the re-evaluation appear to arise from the use of administrative data, longer time horizon and various other refinements such as matching on pre-treatment benefit history. The commissioning of the re-evaluation is to be commended in terms of validation, and the re-evaluation does a comprehensive assessment. The impact evaluations in the combined evaluation and re-evaluation scores Level 3-4 on the Maryland scale, although the bulk of the original evaluation is purely descriptive (unclassifiable, or Level 1-2). Both evaluation and re-evaluation are impaired by the basic design that compares voluntary participants and non-participants, in which matching on observables cannot be taken to imply matching on unobservables.

International comparators

There is a large volume of experimental evidence on similar progammes in the US on mandatory and voluntary programmes aimed at lone parents. Dolton and Smith (2011) provide a comparison with this US literature. Other programmes are listed in Appendix B of Card, David, Jochen Kluve and Andrea Weber (2010) Active Labor Market Policy Evaluation: A Meta-Analysis, Economic Journal, 120(548) F452-F477. International comparisons are also considered in another DWP report Millar and Martin (2003).

Documents examined

Lessof, C., Miller, M., Phillips, M., Pickering, K., Purdon, S. and Hales, J., New Deal For Lone Parents Evaluation: Findings from the Quantitative Survey, (2003a). National Centre for Social Research, DWP Research Report WAE147, March 2003

Millar, Jane and Martin Evans eds. (2003). Estimating the impact of NDLP. Lone parents and employment: International comparisons of what works. J. Millar and M. E. (eds.), Centre for the Analysis of Social Policy - University of Bath, DWP Research Report 181, December 2003.

Dolton, Peter and Smith, Jeffrey A., The Impact of the UK New Deal for Lone Parents on Benefit Receipt (February 1, 2011). IZA Discussion Paper No. 5491

Professor Peter Dolton, João Pedro Azevedo and Professor Jeffrey Smith (2006), The econometric evaluation of the New Deal for Lone Parents, Department for Work and Pensions, Research Report No 356 http://research.dwp.gov.uk/asd/asd5/rports2005-2006/rrep356.pdf

Pathways to Work

Policy objectives

The Pathways to Work package of reforms ('Pathways', for short) is aimed at encouraging employment among people claiming incapacity benefits. Introduced on a pilot basis in three Jobcentre Plus districts in October 2003 and four further districts in April 2004, it requires most new claimants to attend a series of Work Focused Interviews (WFIs). Participants become eligible for increased financial and non-financial support which aims to encourage a move into paid employment. Pathways has been has expanded to cover more districts so that by December 2006 it covered 40 per cent of the country.

Scope of evaluation

Evaluation was carried out by a consortium of research organisations using a range of analytical approaches – both qualitative and quantitative. Evaluation aimed to provide a thorough understanding of the impact of the policy, its net benefit and the experience both of those delivering the intervention and those participating in it. This assessment focusses on the synthesis report, which cites 15 separate sub-reports, and there is at least one other, dating from 2004-2008. Out of these, only 3 provide quantitative impact evaluation, the rest being process evaluation, descriptive statistics about the Pathways population and qualitative discussion. Two quantitative impact evaluations appear to have been carried out in quite close succession – rrep354 (2006) reports on 'early' impacts of the October 2003 and April 2004 pilots on survey respondents. Rrep 435 (2007) reports on impacts of the same pilots, but extending the survey data to 2006 and drawing on administrative data.

Overall methodology

The impact of Pathways was examined using both survey data and administrative data. For a given outcome, comparing the difference between pilot and non-pilot areas before and after the introduction of Pathways gives an estimate of the effect of the programme (difference-in-difference). The early impacts study (rrep354) also used propensity score matching.

Impact evaluation

Two reports, plus CBA, but a small part of the overall evaluation. The aim of the impact analysis was to estimate the effect of Pathways as a whole rather than the effect of a specific component such as the New Deal for Disabled People (NDDP) or Condition Management Programme (CMP). Feeds into CBA.

Policy details

Introduced mandatory Work Focused Interviews (WFIs) to increase the conditionality associated with receipt of incapacity benefits. Introduced a number of innovations for those beginning an incapacity benefits claim, including

- A faster Personal Capability Assessment (PCA);
- A series of Work Focused Interviews (WFIs), mandatory for most customers, carried out by specially-trained advisers;
- A package of new and existing voluntary provision known as 'Choices'. This includes the New Deal for Disabled People (NDDP) and the Condition Management Programme (CMP) – a new programme run in collaboration with local health providers to help individuals manage their disability or health condition;
- A Return to Work Credit (RTWC) for those entering full-time employment;
- In-Work Support (IWS) and other help for those entering employment.

Staggered introduction (2003,2004) in pilot areas followed by national roll-out.

Data

Survey and administrative data. Department for Work and Pensions (DWP) National Benefits Database (NBD) which captures most benefit spells back as far as June 1999. Survey data from two cohorts of individuals making new incapacity benefit enquiry, January-March 2004 (pre-2004-policy) and August-November 2004 (post-2004-policy).

Costs

Covered in detail in a separate report rrep498: staff costs at Jobcentre Plus (the salaries and non-salary expenditures, including travel costs, office expenditures, the rental cost of office space and computer purchases and maintenance, associated with the staff time required to administer the screening tool and to conduct the follow-up WFIs); the costs of the Choices components; payments made to individuals through the Return to Work Credit (RTWC) and the Adviser Discretionary Fund (ADF); costs resulting from fasttracking Personal Capability Assessments (PCAs); and indirect taxes, such as VAT, that result from Government expenditure on Pathways.

Outcome variables

Direct effects on employment, earnings, incapacity benefit health from survey data. Timing of benefit effects from administrative data. Indirect effects on other benefits (not directly triggering Pathways participation e.g. IS, JSA) from administrative data.

Control group

Specially selected areas in which Pathways were not piloted in 2003 and/or 2004. The comparison areas were selected on the basis that they were similar to the pilot areas in terms of economic and social characteristics in the 2001 Census and that Jobcentre Plus had already been introduced (details not provided in the main report, but in a technical report that could not be tracked down). Note potential complications in 2004 due to earlier pilot areas in 2003. Final evaluation uses 2004 pilot areas only. Implicit assumption of no contamination from treatment to control by current or previous pilots, but difficult to assess with no information on location of comparators.

Methodology details

Difference-in-difference: Compare the labour market outcomes for individuals starting new claims in the pilot areas before Pathways was introduced with outcomes for individuals starting new claims in non-pilot areas at the same time and then see how this relationship changes after Pathways was introduced.

Internal validity

The early impacts report presents detail on balancing and explores sensitivity to control variables. The later impact analysis and synthesis report provides less information on these issues. Implications for roll-out of the Pathways nationally on the control group in the longer term analysis are not discussed.

Inference

P-values or graphical confidence intervals reported, but little detail given on methods for constructing these (e.g. clustering).

External validity

Notes potential discrepancy between survey and administrative data, due to impact of Pathways on probability of making a claim. Admin data includes claimants only, whereas survey includes all new inquiries. The representativeness of the survey sample is not assessed in detail, and any issues of non-response bias are not discussed in the impact evaluations. The

impacts assessed are relatively short run, up to 18 months after the pilots introduced, so longer run impacts unknown.

Cost effectiveness

A separate 150 page report is devoted to a comprehensive cost benefit analysis. Takes into account costs and benefits from benefits, taxes, national insurance, administrative and staff costs, earnings. Involves some simulation from the IFS "taxben" tax and benefit system model.

Overall assessment

A long and complex evaluation, with three reports directly related to impact evaluation and cost benefit analysis (plus synthesis report). There is potential overlap with NDDP evaluations, which were based on individual participation in NDDP. The Pathways evaluation is essentially an area based evaluation which looks at the impacts on claimants in pilot areas relative to comparison areas. The impact evaluations are of high quality, and the difference-in-difference design is appropriate, but the Policy Studies Institute evaluations lack of detail on the chosen comparator areas makes assessment difficult (e.g. potential spillovers, displacement). Only the early impacts IFS report provides much detail on the balancing between the treatment and control areas. The synthesis report provides a clear summary of the findings and the net benefits (£ 700-£1600 per incapacity benefit enquiry) The impact evaluation scores 3-4 on the Maryland scale, but impact evaluation is a relatively small part (by volume) of the total.

International comparators

Wide range of active labour market programme evaluations available internationally. Related international disability-specific programmes are described in DWP in house report 90. The evaluation is comparable or better than other recent evaluations of disabiled persons labour market policy internationally, e.g. Human Resources and Skills Development Canada (2010) the Evaluation of the Canada-Manitoba Labour Market Agreement for Persons with Disabilities, Human Resources and Skills Development Canada. Project Network in the US provided a randomised control trial study of ALMP for disabled people in the early 1990s (references provided in Corden 2002).

Documents examined

Adam, Stuart, Carl Emmerson, Christine Frayne and Alissa Goodman Department for Work and Pensions (2006) Early quantitative evidence on the impact of the Pathways to Work pilots, DWP Research Report No 354

Adam, S., Bozio, A., Emmerson, C., Greenberg, D. and Knight, G. (2008) A cost-benefit analysis of Pathways to Work for new and repeat incapacity benefits claimants, DWP Research Report No. 498.

Bewley, H., Dorsett, R. and Haile, G. (2007) The impact of Pathways to Work, DWP Research Report No. 435.

Dorsett, Richard (2008) Pathways to Work for new and repeat incapacity benefits claimants: Evaluation synthesis report, DWP Research Report No 525 http://research.dwp.gov.uk/asd/asd5/rports2007-2008/rrep525.pdf

Work Based Learning for Adults, 2006

Policy objectives

Work-Based Learning for Adults (WBLA) is a voluntary programme designed to help long-term jobless people on a range of benefits move into sustained employment. It offers jobseekers a variety of occupational skills and gives them the opportunity of working towards a recognised qualification that will increase their chances of finding work.

Scope of evaluation

The evaluation follows on from an earlier evaluation of the short run (up to 12 months) impact of WLBA on employment outcomes which found no impacts (Anderson, T., Dorsett, R., Hales, J., Lissenburgh, S., Pires, C. and Smeaton, D. (2004), 'Work-based learning for adults: an evaluation of labour market effects', DWP Report 187, Sheffield: Department for Work and Pensions.). The new evaluation looks at the effect of WBLA on participants who started on the programme during the first quarter of 2002 and follows outcomes up to 40 months, and was intended to cover long run impacts on a wider range of outcomes.

Overall methodology

Propensity score matching of individuals participating in WBLA with other JSA participants, combined with difference-in-difference analysis of the intervention and control groups to adjust for lack of balancing in pre-programme employment levels after propensity score matching. The implementation of the difference-in-difference method is not explained fully, but appears to involve controlling for pre-existing outcome differences between treatment and control groups (p. 23"The difference-in-difference estimator is implemented in a semi-parametric way by including the employment situation before treatment in a regression framework of outcomes (Bergemann et al. 2000, 2005)"). Results are presented graphically, with the per-month intervention effect magnitudes and confidence intervals.

Impact evaluation

Estimation of the effects of participation in one of the WBLA programmes on subsequent benefit claimant status (all benefits, JSA, non-JSA) and employment, up to 40 months after participation. Study find effects from Short Job-Focused Training (SJFT) and Longer

Occupational Training (LOT), but more mixed results for Basic Employability Training (BET).

Policy details

Policy introduced at Job Centre Plus from 2001. Three opportunities of WBLA, Short Job-Focused Training (SJFT), Longer Occupational Training (LOT) and Basic Employability Training (BET). The programme is for jobless people aged 25 and over on Jobseeker's Allowance (JSA) and a range of other benefits, including Incapacity Benefit (IB). Main eligibility is six months or more out of work, but a number of groups, including people with disabilities, can enter the programme earlier. Responsibility for delivery was transferred in 2001 from Training and Enterprise Councils to DWP.

Data

Secondary data from mainly administrative sources. Work and Pension Longitudinal Study (WPLS) combined from data on WBLA participation (source unclear). HMRC employment data from tax records. The data is suited to purpose, although the WPLS data is lacking in individual characteristics which limits the effectiveness of the propensity score matching method (which is reliant on a large set of matching variables).

Costs

Not given.

Outcome variables

Benefit claimant status (all benefits, JSA, non-JSA) and employment status, up to 40 months after participation. These seem appropriate to the programme in question, although others (e.g. wages) could have been considered.

Control group

The comparison group consists of JSA claimants who have not participated in WBLA between the year 2001 and the end of the period of observation in August 2005. JSA claimants are only part of the comparison group if they have been on the JSA register for at least one day between 1 January 2002 and 30 April 2002. 'Potential' WBLA start dates for comparator JSA recipients determined by randomisation. Additionally, all earlier and later participation in alternative programmes of the comparison group are included in the extract, e.g. New Deal for Young People (NDYP) or New Deal for Lone Parents (NDLP). Intervention group is those starting WBLA between January and April 2002. Various other selection criteria used to reduce cross-contamination from other interventions.

Methodology details

The study uses propensity score matching to match JCA recipients who are WBLA participants to JSA recipients who are not. Matching characteristics are age; gender; ethnic group; Jobcentre Plus areas and pre-intervention benefit histories (dummy variables). Because of the limited set of matching characteristics (and presumably because participation on WBLA would be partly determined by worse employment histories) the matched samples are not well balanced on pre-policy employment outcomes (particularly for BET participants). To control for pre-programme differences, the report states that a differences-in-difference estimator is applied. The report is not clear about exactly how this is implemented, or why it is necessary give than the matching variables include variables for the pre-programme benefits history (so participants and non-participants should already be matched by pre-programme outcomes).

Internal validity

There is no evidence presented for testing of sensitivity to alternative specifications. There is evidence of extensive testing for pre-intervention balancing between treatment and control group. The report also notes Ashenfelter's dip, arguing that adjusting treatment and control groups to be balanced on outcomes immediately prior to the programme is inadvisable, because it is common to observe a pre-intervention dip in outcomes for programme participants in studies of interventions of this type. The report claims to address this by using long run (unspecified) pre-intervention outcomes to adjust for treatment and control group differences. However, outcomes immediately prior to the programme appear as matching variables in the propensity score matching process so the efficacy of this procedure is questionable.

Inference

Graphical analysis includes 95% confidence intervals. Tables include standard errors and/or CIs. Standard errors for propensity score matching estimates obtained by bootstrap methods.

External validity

Administrative data, so good external validity, subject to sampling rules. Generalisable to various time intervals using estimated impacts up to 40 months after participation in WBLA programme. Potential displacement issues in that the research design compares JSA recipients on WBLA with matched JSA recipients not on WBLA. It is not completely clear if the measured impacts reflect displacement of non-WBLA participants. This issue potentially important, given both groups could be competing for the same jobs, although the WBLA participation group is small (20,000) relative to the comparator group (800,000) which may mitigate these displacement effects.

Cost effectiveness

There is no statement of the costs of the programme or the monetised benefits, although the outputs could be used to derive monetary benefit estimates.

Overall assessment

Attempts to address many relevant challenges to estimation of the causal impacts of the policy. No randomisation, but provides a matched control group through propensity score matching and adjusts remaining differences between groups using a difference in difference design. Due to voluntary participation, matching on observables does not provide much guidance as to matching on unobservables. As such, ranks as Level 3 on the Maryland scale.

The evaluation presents clear graphical evidence and conclusions on the impacts of the programme from individual level administrative data and attempts to address many of the issues relevant to programme evaluation. Parts of the report are hard to follow in detail, making some aspects of the method difficult to assess.

Potential improvements: greater clarity in describing some aspects of the methods, especially difference-in-difference and final estimation method (the propensity score matching is described in detail). Greater consideration of pre-intervention trends, displacement effects, and more testing of sensitivity of results to alternative specifications (e.g. matching variables, control variables and so on). Study looks only at benefit and employment outcomes, not wages for those who gained employment.

International comparators

There are many evaluations of international work based training schemes which can provide comparators. See online Appendix B of Card, David, Jochen Kluve and Andrea Weber (2010) Active Labor Market Policy Evaluation: A Meta-Analysis, Economic Journal, 120(548) F452-F477

Documents examined

Stefan Speckesser and Helen Bewley (2006) The longer term outcomes of Work-Based Learning for Adults: Evidence from administrative data, Department for Work and Pensions, Research Report No 390 http://research.dwp.gov.uk/asd/asd5/rports2005-2006/rrep390.pdf

Appendix: Evaluation of Business Support Schemes

This appendix provides details of the evaluations considered in the area of business support. The structure of the template was agreed following discussions with the National Audit Office. In completing the templates, for reasons of both feasibility and presentation, we have made use of source material from the original evaluations without any attempt to provide detailed attribution (e.g. through the use of quotes, or the provision of page numbers).

Economic Impact of Business Link Local Service

Policy objectives

Legitimation by Business Link reduces uncertainty for SME managers surrounding the performance of 'hired' consultants; Working with Business Link increases the capacity of SME managers to analyse their problems and derive solutions; The high visibility of Business Link enables SME owner-managers to know where to go to find business advice; The number of consultants dealing with SMEs increases; Once subsidy ends former Business Link clients more likely to seek further external advice

Scope of evaluation

Evaluation assessed the impact of Business Link Local Services on those businesses that received assistance in the 6 month period April to September 2003 and its impact over the subsequent period to May/June 2005. The original aims of the report were:

- Understand and quantify improvements in the performance of the network since the last evaluation in 1988, following reorganisation in 2001;
- Identify examples of good practice and ways of working within the network;
- Update VfM estimates using gross value added (GVA) and other measures;
- Provide a baseline for new arrangements introduced in April 2005 when delivery of Business Link services was transferred to Regional Development Agencies.

Overall methodology

VfM economic impact assessment was built around a methodology which included:

- An extensive telephone survey of approx. 3,500 firms covering BL assisted businesses and a similarly sized control group. Assisted firms sample drawn from population of assisted firms. Survey designed to support an econometric approach to overcome any systematic bias in the type of assisted firms;
- Detailed face-to-face interview survey with 34 firms focusing on those who received 'intensive assistance' to provide more detailed information on the organisational and strategic impact of BL support;
- Interviews with 18 Business Link Organisations (BLOs) and the subsequent development of a detailed typology of alternative brokerage models.

Impact evaluation

OLS plus Heckman selection procedure to correct for selection into treatment.

Policy details

Reforms: Each local franchise became a distinct local body that contracts directly to the Small Business Service (SBS). However, evidence from interlocking directorships suggests only 58.1 per cent of BLOs are really independent; 25.6 per cent have strong interlocking directorships with Chambers of Commerce. The role of the business adviser changed to emphasise brokerage and referral rather than direct help.

Data

Large survey. Given the difficulties with the collection of GVA data, business growth (employment and sales) and sales per employee indicators were used as key performance measures in the econometric models.

Costs

Not clear extent to which had good cost data for localised provision.

Outcome variables

Business growth (employment and sales) and sales per employee indicators.

Control group

Unsupported firms (for the overall impact evaluation); Firms receiving different types of support (for the evaluation of different types of support)

Methodology details

First stage involved development of a series of Probit models of the probability of receiving assistance (reported in Tables 4.1 and 4.2 for intensively-assisted and other-assisted firms respectively). In each case three models are reported with slightly different specifications to give an indication of robustness (Model 3 is the preferred specification). The impact of this assistance is then assessed using OLS models for employment, sales and labour productivity growth. Selection effects (addressed using Heckman selection) are generally weak and take varied signs for intensively-assisted firms. For other-assisted firms results suggest assistance targets better than average firms. Lack of significance on selection variables used to justify focus on OLS models with no selection in main text (selection models are in an appendix). Models compare impact of either intensive assistance versus no assistance (Table 4.3) or other assistance versus no assistance (Table 4.4) and include a wide array of potential determinants of employment, sales, and labour productivity growth. Also report preferred specifications after the removal of extraneous variables with little to no explanatory power. Also model perceived impact of receiving different types of assistance – conditional on

receiving assistance. Additional results then break down by type of brokerage but run similar specifications.

Internal validity

Main issue is way in which deal with selection through Heckmann selection equation. As the report recognises: 'an important issue in operationalising the Heckman type model is the avoidance of too much overlap between the selection and performance models. This is a particular problem in secondary analysis where the variable set may be limited.' Choice of variables shaped by awareness of this problem as well as previous experience of the BL Tracker Study and understanding of the small business literature and the determinants of business growth. In the probit models focus on informational variables and objective and observable characteristics of firms – factors which may have provided the basis for administrative criteria for the targeting of assistance. In outcome models control for more organisational factors and the characteristics of the entrepreneur.

Inference

Very little information provided.

External validity

Some discussion of displacement, but no effect estimated.

Cost effectiveness

Increments to employment growth based on the econometric models are converted into absolute employment gains (between 24,915 and 26,908 jobs). Estimates grossed up to a national scale based on the number of interventions with intensively-assisted firms (n=49,830). Employment impact estimates translated into value added using ratios of value added per employee derived from the ABI (i.e. £27,990 per employee).

Overall assessment

This is one of a number of reports evaluating the impact of Business Support which employ a Heckman selection equation to deal with the selection problem. This is a little difficult to place on the Maryland scale because it uses econometric techniques to try to correct for the fact that the treatment and control groups are not comparable. Although level 4 in principle, given the research design it is unlikely that the comparison group is appropriate (or could be made so through use of the Heckman selection procedure). As a result the report would rate 2 and it would be difficult to improve the Maryland scale rating through a better write up (although this would be desirable, regardless).

This is a slight improvement on some of the other reports (e.g. The two reports evaluation Grant for R&D, SMART and Spurt, discussed below) because there is one paragraph of discussion on the need to have different variables in the selection equation versus the impact

equation. That said, the fact that this receives so little attention makes it very hard to assess the validity of the results. As the report says, the correction for selection makes little difference for intensively assisted firms, but it is impossible to tell whether this is because selection doesn't matter or because the cross-equation exclusion restrictions do not hold (rendering the method of correcting for selection invalid).

In addition, there are three possible categories of assistance – intensive assistance, other assistance, no assistance, but these are modelled as two separate and independent probits (which is a little unusual).

There are a number of ways in which this evaluation could have been improved. In terms of the approach taken, the report could have made much better use of, for example, any area variation in delivery. At a more basic level, the report provides very little information on the analysis. This may reflect the fact that this report (along with many others) is being asked to cover a lot of ground and provide information on so many aspects of the policy and delivery that the impact evaluation becomes 'swamped'. In the circumstances, it may not be surprising that the report's authors do not want to provide additional detail to lengthen an already very lengthy report. But this makes it very hard to form an overall assessment of the robustness of the cost-effectiveness calculations provided.

International comparators

US Manufacturing Extension Partnership (for an overview, see: http://www.ieeeusa.org/policy/eyeonwashington/2011/documents/MEP.pdf)

Documents examined

BERR Economic Impact Study of Business Link Local Service
http://www.bis.gov.uk/files/file40289.doc

Evaluation of Smart (including SPUR) 2001

Policy objectives

A central policy instruments for supporting near market R&D projects by SMEs. The ultimate objectives of Smart are to:

- Promote enterprise and innovation and to increase productivity; and to increase the capacity of SMEs, to grow, invest, develop skills, adopt best practices and exploit opportunities abroad;
- Make the most of the UK's science, engineering and technology base by achieving international excellence and maximising the contribution to the economy and quality of life.

Scope of evaluation

The purpose of the evaluation was to up-date knowledge on the effectiveness and value for money of the scheme. The evaluation identifies 15 questions that it attempts to answer to help meet that broad objective.

Overall methodology

Based mainly on a survey of 513 firms that received Smart awards between 1988 and 1998 and a comparison group survey of 191 firms that applied unsuccessfully for an award (covering 1995-1998 only because details for unsuccessful applicants in earlier years not available).The main components of the methodology were:

- Analysis of data to identify patterns and trends in applications and awards;
- Exploratory interviews (20 in total) with scheme administrators;
- A survey of recipients of Micro-project, Technology Review and Technology Study awards (involving telephone interviews with 29 firms);
- A survey of Smart grant recipients (telephone interviews with 468 firms selected to be representative of the wider population of award winners in terms of the type of award received, the size of award recipient and period of award);
- In-depth interviews with Smart grant recipients (involving 45 face-to- face interviews with a representative sub-sample of award winners, coupled with an independent examination of files relating to some of their projects);
- A survey of unsuccessful applicants for Smart grants (involving 191 telephone interviews with firms selected to be representative of the wider population of non-award winners).

Impact evaluation

Multivariate analysis comparing outcomes for Smart/GRD award winners to unsuccessful applicants. OLS plus Heckman selection procedure to correct for selection into treatment.

Policy details

Piloted in 1986 and was fully implemented in 1988; provided finance of more than £200 million to more than 3,000 companies. Smart support entailed provision of grants to enable companies with fewer than 50 employees to undertake Feasibility Studies to research the technical feasibility of concepts. The scheme also provided grants to enable all companies with fewer than 250 employees to undertake Development Projects to work-up concepts to pre-production prototype stage. Design and delivery of the scheme have changed over the years. 1991: SPUR introduced (for companies with up to 500 employees) 1994: Eligibility for SPUR restricted to companies with fewer than 250 employees. In 1997 there was a general rationalisation of DTI schemes resulting in SMART, SPUR, SPURplus and RIN being incorporated into a single scheme, Smart. 1988-1998 Applications: 14,770; Awards 4034; Unsuccessful 10,736. Broad distribution across sectors, although some sectors more represented than others. 66% of recipients claim that SMART fully additional; 32% partly additional (timing, scope, scale); 2% non-additional. 50% of non-funded went ahead anyhow (to get consistent with funded additionality figures suggest that ‘perhaps they were rejected for being non-additional’; although not given as a reason when asked and these firms are younger).

(although details on these only available towards end of the period 1995-1998).

Data

Large amount of data on firms collected from detailed survey and interviews with firms (although data on characteristics limited or not exploited in impact evaluation).

Costs

Data on gross spend by year is available.

Outcome variables

Large variety of outcome variables tackled in the surveys. Performance measures used for the impact analysis were: turnover, employment, productivity (the ratio of turnover to employment), and exports (but NOT R&D which is available)

Control group

For the impact analysis, control group is 191 firms that applied unsuccessfully for an award

Methodology details

Multivariate analysis used to examine the range of determinants of business performance and, in particular, the contribution of Smart/GRD. Purpose is to test whether, and to what extent, observed differences in business performance as between Smart/GRD award winners and unsuccessful applicants are attributable to Smart/GRD itself; or whether, and to what extent, the differences are attributable to other variables. Estimation for the sample of successful and unsuccessful applicants a separate impact analysis for each year 1995 to 1998 for each of turnover, employment, productivity (the ratio of turnover to employment), and exports. In each case a parsimonious specification based on the Law of Proportionate Effect (LPE) was adopted, in which closing year performance is a function of opening year performance. In addition to opening year values of the relevant performance variable, explanatory variables also include the age of the applicant, and (as dummy variables) the growth objectives of the applicant at the time of applying to participate, and a dummy variable representing success in applying for an award. Year-by-year levels regressions of outcomes on limited set of firm characteristics plus treatment dummy for sample of successful and unsuccessful firms. Only coefficients reported are those for coefficient of interest in one set of by-year specifications (table 5.3). No detailed results reported. Table 5.3 reports results on dummy variable for award. 16 coefficients (outcomes by year). Nothing significant at the 10% level.

Internal validity

To correct for selection bias the evaluation employs the Heckman selection model. First stage requires the estimation, for each year, of a probit equation determining the factors which distinguish the successful from the unsuccessful applicants. This then used to construct a variable which, when included alongside the other independent variables in the performance equations, 'corrects' for selection bias. Finds a strong selection bias into the successful application group based on size in years 1995, 1996 and 1998 (award winners bigger) and this was, therefore, corrected for in those years.

Inference

Some discussion in the report of the fact that 'significance tests corrected for heteroscedasticity'. Some discussion of goodness of fit. Report suggests that 'appropriate diagnostic tests revealed that the equations are free from omitted variable bias, with the exception of the productivity equations, where there were signs of mis-specification in one or more years.' No further detail provided.

External validity

There is very little informal discussion (and no formal discussion) of external validity. Displacement (e.g. where would sales go if you stopped trading) dealt with through beneficiary surveys. Doesn't try to assess multipliers (suggests not so applicable for these type of schemes).

Cost effectiveness

Net outputs calculated as Gross effects minus self-reported deadweight minus self-reported displacement. Control totals are used to gross-up from per firm sample estimations to whole-scheme estimates of impact. Between 1988/1989 (the first full year of the schemes operation) and 1999/2000 (the latest financial year for which complete data are available) Smart cost £230.5 million (Table 7.4). Get value for money ratios by dividing one by the other.

Overall assessment

This is one of a number of reports evaluating the impact of Business Support which employ a Heckman selection equation to deal with the selection problem. As with the other reports using similar techniques, this is a little difficult to place on the Maryland scale because it uses econometric techniques to try to correct for the fact that the treatment and control groups are not comparable. Although level 4 in principle, given the research design it is unlikely that the comparison group is appropriate (or could be made so through use of the Heckman selection procedure). As a result the report would rate 2 and it would be difficult to improve the Maryland scale rating through a better write up (although this would be desirable, regardless).

This is a frustrating report. The authors have quite good data available to them – (fairly extensive information on characteristics and outcomes for a sample of successful and

unsuccessful firms). As is recognised in the report, the unsuccessful firms may not form a perfect control group. The authors try to deal with this using a Heckman selection procedure which is hard to understand. Specifically, it appears that size is used as the firm characteristic that determines selection – but size almost certainly directly affects outcomes as well (rendering the cross-equation exclusion restrictions, and thus the correction for selection, invalid).

Assuming that this procedure imperfectly deals with the issue of selection, it seems reasonable to think that this would bias upwards the coefficient on outcome variables (assuming that successful firms are perceived as 'better bets' for some reason unobserved by the researcher). It's surprising, then, that the estimates for outcome variables find essentially no significant impact of receiving an award (one or two coefficients are marginally significant, although this is not that surprising when looking at 16 coefficients at 10% significance levels). The report points out that it is 'only possible in a statistical analysis to control for only a limited range of variables which may affect performance.' This is certainly true, although it does not explain the findings if the assumption of upward bias is correct.

It is very difficult to consider why this is happening, or the robustness of this lack of impact, because the report presents too little information on the statistical analysis taking place (no full results are reported, only coefficients and standard errors for the coefficient of interest; there are no robustness checks; there is very limited discussion of the results). Instead, the report turns to subjective (i.e. self-reported) assessments based on award winners' own views on how Smart had affected aspects of their performance, and on what would have happened without the scheme. This produces *wildly* different results – suggesting strong additionality and little deadweight. Nothing is done to reconcile these contradictory findings and the self-reported additionality is then taken at face value and used in the rest of the report.

Unsuccessful applicants were asked about whether projects went ahead anyhow, and benchmarking against successful applicants produces contradictory findings (many unsuccessful projects go ahead). The report suggests that this might be because the applications were rejected precisely because of concerns over additionality. But this isn't explicitly considered, is not given as a reason as to why applicants think they were rejected and seems puzzling as unsuccessful firms generally much smaller (so we might expect additionality to be bigger for the unsuccessful group).

International comparators

Einio, E. (2011) The Effects of Government R&D Subsidies on Company Performance: Evidence from the ERDF Population-Density Rule [and references therein].

Documents examined

Evaluation of Smart (including SPUR) 2001: Final Report
http://www.bis.gov.uk/files/file22000.pdf

Evaluation of Grant for Research and Development & Smart

Policy objectives

Grant for Research and Development (GRD) set up to help:

- Increase business spend on innovation, including R&D;
- Increase in the proportion of firms that innovate;
- Increased take-up by UK business of the new technology created by the R&D.

Intermediate objectives:

- Increase the productivity and profitability of assisted SMEs;
- Increase and improve technology use and adaptation, and research and development by individuals and SMEs to improve the overall innovation performance of the SME sector;
- Increase the number of successful high growth firms that achieve their potential and to contribute to an enterprise climate that encourages investment in innovative technology by individuals, firms and financial institutions.

Scheme's longer-term objectives:

- To overcome the reluctance of SMEs to undertake risky R&D by sharing costs and the risks associated with projects, and to foster a recognition of the importance of maintaining an ongoing programme of R&D;
- To encourage others to invest in potentially risky technological R&D through the knowledge that RDAs have appraised the financial and technical aspects of a project and is prepared to invest public money;
- To support firms to prove technical and commercial feasibility (Research / Feasibility projects) and to develop prototypes (Development projects).

Scope of evaluation

The main aim was to assess the achievements and impact of GRD, and its predecessor Smart, on the national economy (during the period from the last evaluation in 2001 – discussed above – and covering operation of scheme from 01/04/98 to 31/03/08).

Overall methodology

The main components of the research programme were:

- Survey of GRD recipients (telephone interview with 659 businesses)
- Follow-up interviews with 40 Smart / GRD grant recipients.
- Survey of unsuccessful applicants for Smart / GRD grants.
- Stakeholder Survey with almost one hundred organisations in RDA areas

Impact evaluation

Multivariate analysis comparing outcomes for Smart/GRD award winners to unsuccessful applicants. OLS plus Heckman selection procedure to correct for selection into treatment.

Policy details

The GRD scheme was introduced by DTI on 1 June 2003 as a replacement for the former Smart scheme. Since its introduction in April 2003, GRD has helped almost 1,700 SMEs to research and develop technologically innovative new products and processes through over £130Mof grant funding.

Data

Large amount of data on firms collected as part of detailed survey and interviews with firms (although data on characteristics either limited or not exploited in impact evaluation).

Costs

Data on gross spend by year is available.

Outcome variables

Firm size is measured in terms of turnover and employment. Large variety of outcome variables tackled in the surveys. Performance measures used for the impact analysis were turnover and employment. No R&D (similar to 2001 report, see below) but also no productivity (in contrast to 2001 report, see below).

Control group

For the impact analysis, control group is a sample of firms that applied unsuccessfully for an award.

Methodology details

See above (evaluation of Smart – including SPUR 2001). Data is pooled across years, otherwise appears identical to 2001 analysis. Only coefficients reported are those for coefficient of interest in one set of by-year specifications in table 5.3. No detailed results reported. Table 5.1 reports results on dummy variable for award on base year size and on 'ambitions'. Basic OLS significant, but nothing once use (essentially undocumented) Heckman procedure to address selection problems.

Internal validity

Appears to use exactly the same Heckman selection model as 2001 report (see above).

Inference

As with the 2001 report, some discussion of the fact that 'significance tests corrected for heteroscedasticity'. Some discussion of goodness of fit and diagnostic tests. No further detail provided.

External validity

There is very little informal discussion (and no formal discussion) of external validity. Displacement (e.g. where would sales go if firm not supported) dealt with through beneficiary surveys. In contrast to 2001 report (see above) some attempt to get at multipliers through beneficiary surveys (questions on purchases of firms and where workers live)

Cost effectiveness

Net outputs calculated as Gross effects minus self-reported deadweight minus self-reported displacement. In contrast to 2001 report (see above) also uses self-reported multipliers and average duration to get at cumulative net effect. Control totals are used to gross-up from per firm sample estimates to whole-scheme estimates of impact. Between 1988/1989 (the first full year of the scheme's operation) and 1999/2000 (the latest financial year for which complete data are available) Smart cost £230.5 million. Get value for money ratios by dividing one by the other. In contrast to 2001 report provides GVA impacts in addition to employment and turnover.

Overall assessment

Overall evaluation of this report is as for the earlier 2001 evaluation (discussed above). Although, if anything, this report provides less detail – e.g. on the selection equation – than the 2001 evaluation. As with that report, this is a little difficult to place on the Maryland scale because it uses econometric techniques to try to correct for the fact that the treatment and control groups are not comparable. Although level 4 in principle, given the research design it is unlikely that the comparison group is appropriate (or could be made so through use of the Heckman selection procedure). As a result the report would rate 2 and it would be difficult to improve the Maryland scale rating through a better write up (although this would be desirable, regardless).

Once again, assuming that the Heckman procedure imperfectly deals with the issue of selection, it seems reasonable to think that this would bias upwards the coefficient on outcome variables (assuming that successful firms are perceived as 'better bets' for some reason unobserved by the evaluation team). This happens for the OLS estimates in this report (which pool by year) further highlighting the fact that it is surprising that this didn't happen with the 2001 report discussed above. As with the 2001 report, the lack of significance in the impact evaluation is essentially ignored, for similar reasons (it is 'only possible in a statistical analysis to control for only a limited range of variables which may affect performance.') Once again, it is very difficult to understand why self-reported additionality is so different from the impact evaluation. Similarly it is difficult to consider the robustness of this lack of impact because the report presents too little information on the statistical analysis taking place (no full results are reported, only coefficients and standard errors for the coefficient of interest; there are no robustness checks; there is very limited discussion of the results).

Instead, as before, the report turns to subjective assessments based on award winners' own views on how Smart had affected aspects of their performance, and on what would have happened without the scheme). Once again, this produces *wildly* different results – suggesting strong additionality and little deadweight. As with the 2001 report, nothing is done to reconcile these contradictory findings and the self-reported additionality is then taken at face value and used in the rest of the report. It appears that unsuccessful applicants were asked

about whether projects went ahead anyhow, but there appears to be no attempt to use this to benchmark the self-assessed additionality of recipients.

Taking the two reports together, it's hard to avoid reaching the conclusion that the impact evaluation reports may have been ignored because they suggested that the scheme delivered no additionality.

International comparators

Einio, E. (2011) The Effects of Government R&D Subsidies on Company Performance: Evidence from the ERDF Population-Density Rule [and references therein].

Documents examined

Evaluation of Grant for Research and Development and Smart 2009
http://www.bis.gov.uk/files/file52026.pdf

Evaluation of the Manufacturing Advisory Service 2007

Policy objectives

The MAS is a significant Government intervention supporting the manufacturing sector and in particular small and medium-sized enterprises.

Scope of evaluation

The scheme was launched in 2002 and this evaluation was conducted to provide an independent review of the achievements of the MAS, its effectiveness and impact in the first three years of its operation (2002–2005), and to make recommendations to inform policy and delivery of this intervention in the future.

Overall methodology

The evaluation focused on the most significant MAS support packages (Level 2 diagnostic and Level 4 consultancy). A combination of qualitative and quantitative research methods was deployed. These included:

- Interviews with key staff in the Regional Development Agencies (RDAs) and the MAS regional centres;
- A telephone survey of 946 firms that received Level 2 and/or Level 4 assistance from the MAS between June 2002 and June 2005;
- A survey of a control group of 401 firms that did not receive MAS support
- Case studies with 20 beneficiary companies that received Level 4 support

- Econometric analysis to identify the characteristics of MAS users and explore attribution and impact of the MAS intervention in quantitative terms.

Impact evaluation

Five main types of analysis. Three of these use probit models to look at the extent to which firm characteristics affect take up of different types of assistance. The other two use OLS to look at the impact of MAS assistance on firm outcomes and self-reported organisational aspects of the firm (e.g. use of equipment).

Policy details

The MAS was established and launched by the DTI in partnership with the RDAs in 2002. The rationale for MAS was essentially about providing 'practical hands-on assistance from experts to enable firms to adopt new methods, processes and technologies to improve their productivity and quality performance, and ultimately improve their competitiveness'. Although the scheme serves manufacturing businesses of any size, the focus is on small and medium-sized (SME) manufacturers.

Data

Large amount of data on firms collected as part of detailed survey and interviews.

Costs

Data on gross spend by year and by type of intervention is available.

Outcome variables

Employment, turnover, productivity and GVA percentage change from surveys of participants and non-participants.

Control group

A sample of firms that did not receive MAS support.

Methodology details

Econometric analysis i.e. probit analysis, to determine the key characteristics of the MAS-users. Five stages:

1. Probit analysis to determine whether firm, market and owner-manager characteristics make firms more or less likely to take up MAS of any kind.
2. A probit analysis to determine whether firm, market and owner-manager characteristics make firms more or less likely to take up the MAS level 4 assistance - using data on both MAS users and non-users.

3. A probit analysis to determine whether firm, market and owner-manager characteristics make firms more or less likely to take up the MAS level 4 assistance - using data on only MAS users.
4. Multivariate analysis that investigates how the level of MAS support and/or days of assistance impact on employment, turnover, productivity and GVA percentage change, controlling for firm, market and owner-manager characteristics. This is calculated from the respondents' answers on firm performance related to 2001 and to the present (March/April 2006).
5. Multivariate analyses that focuses on the respondents' own opinions on the success of the MAS assistance with regards to delivery improvement, more efficient use of equipment, increased GVA, improved just-in-time manufacturing processes, improved reduction of scrap and percentage change in turnover.

The core set of variables used in the probit and multivariate analyses are: Size of firm – measured by the number of employees in 2001; Region in which the firm is located; Sector in which the firm operates; Whether the firm exports; Qualification of directors/owners; The number of years in business; Firm's position in supply chain; Whether or not the firm recorded profits in 2001; The extent of competition in the markets in which the firm operates; The results for changes to employment, turnover, GVA and productivity indicate that the level of MAS intervention is not a statistically significant determinant of employment, GVA, productivity and turnover.

Internal validity

There does not appear to have been any attempt to correct for selection bias in to treatment (even though the first three stages of analysis report that firm characteristics affect take up of MAS services).

Inference

No discussion (in main report or appendix)

External validity

No discussion (in main report or appendix)

Cost effectiveness

(1) Average benefit per intervention (weighted to take into account additionality and net of deadweight, displacement and taking into account the counterfactual – all from self reported); (2) Estimated number of interventions with quantifiable benefits (3) Estimated total benefit for assisted population with quantifiable benefits (4) Estimated public funding on Level 4 assisted firms (60% - 80%) (5) Value for Money (VfM)/estimated return on public funding allocated to the MAS Level 4 (6) Implied Annual Internal Rate of Return over a 5 year period

Overall assessment

Entry in to the scheme is voluntary so it is difficult to be sure that the control group (firms not supported by MAS) is valid. Indeed, the report itself shows that firm characteristics impact take up of MAS (in the first three stages of analysis). A robust impact analysis would need to address this issue, but the evaluation did not do this (and the issue is not even discussed in the technical appendix to the report). This means that, as it stands, the report rates as level 2 on the Maryland scale. If the report did more to demonstrate comparability, this might be increased to a solid level 3 (although this seems unlikely given the problems of selection in to treatment are likely to render the untreated as an invalid comparison group which means this is unlikely to be improved above a level 2 on the scale).

Assuming that there is selection, in contrast to some other evaluations (e.g. of GRD, SMART and SPUR considered above) it is very hard to be sure about the direction of bias. Do better or worse firms end up taking part in MAS? If better firms take part in MAS results will be downward biased and vice-versa.

It is debatable whether any econometric analysis based purely on comparison of participants to non-participants could effectively address this problem. This suggests that successful evaluation probably needed some element of randomisation to have been embedded in the project delivery (e.g. in terms of identifying clients). It is possible that the evaluation could have addressed some of these problems by using variation across regions as the he MAS regional centres differ in many ways. Particularly, as the centres appear to vary in terms of the way they run their business, their approach to 'recruiting' and targeting businesses, their funding streams and their partnerships. This may have provided some source of exogenous variation to allow identification of the impact of policy. Regardless, it is surprising that the report gives essentially no consideration to these problems.

As with the evaluations of GRD, SMART and SPUR, the cost-effectiveness calculations are based on self-reported assessments of additionality because the econometric analysis is 'inconclusive in confirming whether firms who receive MAS support over the period perform any better as a result of the support provided than those manufacturing firms that do not'. The report argues that 'this is partly due to statistical factors and partly due to the strong likelihood that it will be early days for the full effects of the MAS intervention to have been felt, particularly by the considerable number of firms who received MAS assistance later in the study period.' While this may be true, it is hard to understand why this circumvents the problem of the fact that it is early days for the impact to be felt. The report notes that 80% of firms said that the benefits would be felt over at least 5 years, which implies that we may not have seen all of the benefits yet. But some of the benefits should be apparent and thus potentially identifiable in the econometric analysis.

It is also possible, of course, that the econometric analysis is correctly identifying the fact that MAS has no impact on firms (ignoring the problems of selection discussed above). This is highly problematic because, once again, self-assessment produces *wildly* different results – suggesting strong additionality and little deadweight. As with other reports already

considered, nothing is done to reconcile these contradictory findings and the self-reported additionality is then taken at face value and used in the evaluation of cost-effectiveness.

International comparators

Manufacturing Extension Partnership in the US

Documents examined

Evaluation of the Manufacturing Advisory Service: Main Report
http://www.berr.gov.uk/files/file38877.pdf

Impact of the Manufacturing Advisory Service: Annexes A-F
http://www.bis.gov.uk/files/file38878.pdf

Evaluation of Regional Selective Assistance 1991-1995

Policy objectives

During the time period covered by the evaluation RSA was the main British scheme of financial assistance to industry. It provided discretionary grants to companies creating or safeguarding employment in the Assisted Areas (AAs) of Great Britain.

Scope of evaluation

Main objectives of this evaluation were to measure the effectiveness of RSA in terms of employment generation in Assisted Areas and to measure the cost of RSA grant payments in the period 1991-1995 in terms of the net cost per net job. Specific issues:

- whether RSA was meeting its objectives and whether there are alternative ways of meeting these objectives;
- whether consideration should be given to more targeted approaches to different types of project;
- the relative cost effectiveness of internationally mobile projects compared with domestic expansion/reinvestment;
- the relative cost effectiveness of large versus small projects;
- whether the criteria applied when jobs were safeguarded led to genuinely competitive businesses being assisted;
- the impact of extending RSA in August 1993 to previously ineligible areas;
- the approach to product market displacement in the appraisal process.

Overall methodology

The methodology adopted for this evaluation broadly followed that used in earlier RSA evaluations. It combined three main approaches:

- the analysis of data held on the Departments' Selective Assistance Management Information System (SAMIS) for projects offered RSA in the calendar years 1991- June 1995 inclusive.
- A survey of projects from a sample of those which had been completed by mid 1998, by questionnaire and interview;
- discussions with case officers and regional development organisations.

Impact evaluation

The study brief required that the consultants follow broadly the methods used in two previous evaluations of RSA. This method estimates policy efficiency and cost effectiveness principally by using a measure of net cost per net job. This is based on adjustments to the gross employment information provided by firms and recorded on the SAMIS database and a number of adjustments to the gross grant paid.

Policy details

Applicants can be companies, partnerships or sole traders. Assistance is provided to establish a new project or expand/modernise an existing business, to set up research and development facilities, or to take the next step from development to production. Assistance is not however available for transferring existing plant from one part of the country to another. To be eligible for RSA, projects must have a net positive impact on AA employment levels; create or safeguard jobs; be viable; contribute positively to the national economy; and need grant support to take place. It is also necessary to show that the investment would not proceed without grant.

Data

SAMIS plus survey means good data on participants but no-data on non-participants.

Costs

Good data is available on the costs of projects from SAMIS. The cost-effectiveness assessment makes a number of adjustments on the costs side. These include:

- deduction of RSA assistance returned to the exchequer in the form of taxes;
- the conversion of grants to a common constant price;
- discounting to take account of the timing of grant payments.

Outcome variables

Employment

Control group

The evaluation did not use a control group.

Methodology details

The report relies on self-reported additionality and displacement.

Internal validity

Not applicable.

Inference

According to the report: ‘Data is generally presented for both the sample and population, the latter based on grossed up estimates. The results are the average in each case, e.g. the average net cost per net job. However, it is important to know how much the averages might have

varied if a different set of projects in each category had been sampled. This potential range of variation is presented in the form of ‘confidence intervals’ around the averages, which are calculated to contain most of the potential averages that may have arisen from alternative samples. These are calculated using standard statistical procedures’. No further detail is provided.

External validity

Not discussed.

Cost effectiveness

Adjustments to employment outputs are made to take account of several factors:

- How long jobs created or safeguarded by RSA projects were expected to last;
- What would have happened to jobs in the absence of RSA - 'additionality';
- Displacement effects of RSA projects on competitor firms in the local or other Assisted Areas;
- The inter-industry linkage effects of RSA projects;
- The macro-economic feedbacks associated with RSA projects.

The sample was disaggregated into 8 subgroups by size of grant and whether the project was creating or safeguarding jobs. Adjustment factors estimated for each of these different groups were then used as a basis for grossing up.

Overall assessment

This report rates level 1 on the Maryland scale. There is no random assignment and no control group is used.

This study is particularly interesting in the context of our review of evaluations because of the fact that the brief explicitly restricted the approach to that used by previous studies. This resulted in the study only collecting data on participants (despite the fact that other evaluations in the department, undertaken during a similar time period, had already begun to collect data on comparators who did *not* get assistance).

The report is also interesting for the fact that the consideration of alternative approaches does not raise the possibility of identifying a control group and constructing the counterfactual. Instead, it suggests that the main alternative to the survey approach used in this study would involve the systematic modelling of the intra-regional and inter-regional effects of regional policy. In essence the basis of such a model would either be input-output analysis or dynamic input-output methods which extend to Computable General Equilibrium (CGE) approaches. Later evaluations of RSA, including academic studies, show that this is *not* the case. SAMIS provides good data on both successful and non-successful applicants so arguably RSA is one of the business support schemes that would be most amenable to proper impact assessment using modern programme evaluation techniques.

More narrowly, in terms of the approach adopted, there are some concerns. In particular, there is very little info on how self-reported additionality, etc were grossed up and how confidence intervals were calculated. The decision to gross up based on eight different sample sizes means that the group means used were based on very small numbers of firms (approximately 20 per group). Good practice would involve much more discussion of the sensitivity to the decision on group sizes. It's interesting to note that this willingness to identify several additionality coefficients for different groups is not limited to the RSA

evaluation. The evaluation of SRB, undertaken around the same time, reported around 60 different additionality coefficients based on a sample size of less than 200. In turn, these additionality coefficients then formed the basis for much of the evaluation work underpinning the evaluation of the effectiveness on RDAs showing how problems are easily transferred across seemingly unrelated evaluations.

International comparators

Evaluations of Germany's Gemeinschaftsaufgabe `Verbesserung der regionalen Wirtschaftsstruktur"'. See, for example, Steinwender, C. (2010) Job Creation Subsidies and Employment. Empirical Evidence for Germany (and references therein)

Documents examined

Evaluation of Regional Selective Assistance 1991-1995
http://www.bis.gov.uk/files/file22008.pdf

Regional Selective Assistance and Selective Finance for Investment in England

Policy objectives

RSA Scheme was a prominent feature of regional policy in Great Britain for more than 30 years (1972-2004) and was used to address labour market inequalities. RSA was replaced by the SFIE Scheme in April 2004 with a focus on increasing productivity and the proportion of skilled jobs in Assisted Areas of England.

Scope of evaluation

Research objectives set out by BERR were to:

- Test the validity of the key assumptions underlying the rationale for the old RSA Scheme and the new SFIE Scheme;
- Assess the outcomes of funded projects against objectives with the key measure being productivity, skilled jobs and spillovers; in the case of the RSA Scheme the principle objective in the period 2000-04 was to increase jobs.

Report contains literature/theory review, descriptive evidence on program participants and non-participants, econometric evaluation of impacts on employment growth, qualitative evidence on self-reported benefits and experiences of operation.

Overall methodology

Variety of methods used including face to face interviews and a lot of descriptive statistics, but core of the evaluation methodology is comparison of beneficiary and non-beneficiary

firms using instrumental variables or Heckman selection correction to adjust for selection bias. SFIE evaluation is case study only. Qualitative evidence from beneficiaries on self-reported 'additionality'.

Impact evaluation

Impact of the RSA financial assistance 2002-04 on employment growth 2004-2006.

Policy details

Financial support provided to business in "Assisted Areas" under the RSA Scheme (£462.5 million offered to 784 businesses) in the period 2000-2004, and its replacement SFIE Scheme since April 2004 (£100.1 million offered to 526 businesses). Details of rules under which firms were allocated support are not provided, but were handled by RDAs based on applications for grant assistance.

Data

The econometric analysis is based on a bespoke survey of around 700 RSA assisted and non-assisted businesses in England. Response rates are not high: 60% for beneficiaries surveyed and only 20% for non-beneficiaries. Sampling frame was beneficiaries since 2000, but sampling frame for non-beneficiaries unspecified.

Costs

Details on subsidies provided, but no information on other administrative costs.

Outcome variables

Employment growth 2002-2004 is the only impact investigated in the econometric evaluation. Qualitative/desciptive evidence provided on a range of other things.

Control group

For impact evaluation, comparator group of non-beneficiaries (only 20% response rate). No details on how these were sampled (e.g. were they in assisted areas?). Report shows that the beneficiary and non-beneficiery respondents were not balanced on a number of dimensions.

Methodology details

Core methods in evaluation are either instrumental variables or Heckman selection correction (control function). The text is quite vague about the choice of instruments; appendix A reveals that the instrument for RSA beneficiary is the existence of a published business plan, plus age. Details on exclusions from Heckman selection term are vague. Additional results presented for effect of size of grant on employment growth amongst firms receiving the

grant, with Heckman sample selection adjustment. There are many additional claims based on survey and (10) case studies which make no use of the non-beneficiary control group.

Internal validity

Report clearly demonstrates that the beneficiary and non-beneficiary samples are unbalanced and that there is selection into RSA. Solution is IV and Heckman selection correction. One instrument is public business plan, which it is argued helps determine RSA, but is not correlated with employment growth, and yet this has no predictive power in first stage RSA equation. Firm age is an additional instrument, but this has no theoretical justification and appears to have been chosen based on lack of statistical correlation between age and employment growth. Implementation of the selection correction method is vague, with seemingly arbitrary decisions about what to include in the selection correction equation and what to exclude from the employment growth equation (appendix A notes the "avoidance of too much overlap between the selection and performance models"). In the estimates of the effect of the size of the grant, the possibility that the size of grant is potentially affected by unobserved firm characteristics is ignored. Qualitative evidence makes no attempt to establish that the self-reported outcomes were attributable to the policy.

Inference

Tabulated regression results report t-statistics, but no details given on any clustering assumptions.

External validity

This is not discussed explicitly. The sample is small, and, necessarily, a very selected group of firms. The evidence could not therefore be used to generalise to the effect of financial assistance generally, but only to firms of the type selected for this kind of assistance. Appears to be no discussion of issues of non-response bias, and estimates do not appear to have been weighted to account for this. Displacement issues are discussed, and some (inconclusive, unclear) descriptive evidence presented on extent of relocations and within-region competition. Limited time period for effects to be realised is noted.

Cost effectiveness

Costs are discussed, but no cost effectiveness or cost benefit calculation.

Overall assessment

This is a mixed-methods evaluation, with one chapter specifically on the impact evaluation. There is an attempt to construct a counterfactual from a control group of non-beneficiaries, although details of the survey design are absent. Econometric methods used for evaluation are potentially appropriate, although general implementation and discussion are poor. As it stands, this report would rate as level 2 on the Maryland scale because it does not do enough

to demonstrate that the econometric techniques deal with the lack of comparability of the control and treatment groups. If the report did more to demonstrate comparability, this might be increased to a solid level 3 (although this seems unlikely given the problems of selection in to treatment are likely to render the untreated as an invalid comparison group which means this is unlikely to be improved above a level 2 on the scale).

The choice of instruments in the IV analysis is not well defended – one of the key instruments does not predict RSA assistance in the full sample - and the exact implementation of the Heckman selection correction approach is not fully explained. Very few tests of the identifying assumptions (e.g. predictive power of instruments, correlation of instruments with pre-treatment characteristics). This is one program where alternative quasi-experimental approaches might have been available given the potential eligibility rules, and even propensity score matching on characteristics related to eligibility might have been worthwhile. Much of the report (and all the evidence on SFIE) are of dubious value in determining the impacts of the programme.

International comparators

See above (Evaluation of Regional Selective Assistance 1991-1995)

Documents examined

Hart, Mark Hart, Nigel Driffield, Stephen Roper, Kevin Mole (2008) Evaluation of Regional Selective Assistance (RSA) and its successor, Selective Finance for Investment in England (SFIE), BERR Occasional Paper 2

Economic Evaluation of the Small Firms Loan Guarantee Scheme

Policy objectives

The SFLG was the government's primary debt finance instrument, established in 1981. In January 2009, SFLG was replaced by the Enterprise Finance Guarantee. SFLG addressed market failure in provision of debt finance by providing a Government guarantee to banks in cases where a business with a viable plan is unable to raise finance because they cannot offer security for their debt or lack a track record.

Scope of evaluation

The main objective of this research was to provide a comprehensive assessment of the wider economic impact of SFLG arising from supported businesses being able to access loans that they would otherwise not have received. The specific objective of this evaluation was to assess the impact of SFLG on a number of business outcomes and through a Cost-Benefit Analysis, determine whether the scheme was cost effective to the economy. In particular, the

evaluation focuses on the impact of SFLG on business growth, labour productivity, and propensity to introduce new technology and innovation and also market internationalisation.

Overall methodology

The research uses a comparison group methodology to assess the counterfactual. The counterfactual was established by constructing a matched sample to compare the performance outcomes of those accessing SFLG supported loan as against a sample of similar businesses not accessing SFLG loans.

Impact evaluation

Evaluation uses businesses self-reported assessment of business performance and scheme impact. Telephone interviews were conducted businesses who had received an SFLG loan in 2006, alongside a matched sample of non-users from the general business population. The comparison sample group was matched to the SFLG group in terms of company legal status and broad industry sector (to one level SIC). In total, 1,488 businesses were surveyed including 441 SFLG supported businesses and 1,047 unassisted businesses.

To identify impact of SFLG, used matching on sector, age and initial size of businesses to control for key differences in characteristics between the sample groups.

Policy details

SFLG first established in 1981 as Government's principal debt finance instrument supporing access to finance for small businesses. Around 4,500 businesses supported per year. Guarantee covers up to 75% of qualifying loans of amounts up to £250,000. In return for the guarantee, the business pays BIS an annual premium of two per cent of the outstanding balance of the loan, assessed and paid quarterly. Businesses do not apply for SFLG directly. SFLG operates as a tool for the lender to use at their discretion alongside their normal commercial lending practices (and is not designed to replace mainstream lending decisions). However, SFLG is often used as part of an overall package of finance that borrowers put together. It is estimated that SFLG accounts for roughly 1% of all SME lending by value.

Data

Large amount of data collected from detailed survey and interviews with firms

Costs

Report provides estimates of net costs to the Exchequer (costs of called in guarantees plus administration costs less premium income) although only based on loans of 1.5 to 2.5 years duration.

Outcome variables

The impact of SFLG is assessed on a number of business outcomes including employment change, sales change, labour productivity, likelihood to export, propensity to introduce new products and processes.

Control group

Group of non-SFLG recipients from the general business population.

Methodology details

Survey collected information on additionality including finance deadweight and market displacement amongst SFLG supported businesses as well as growth orientation, employment and sales growth, product and process innovation, labour market history of the owner, geographic market focus and internationalisation.

When assessing finance additionality SFLG recipient group is compared to firms who received a conventional bank loan. No statistical difference between SFLG and comparison group viewed as a positive outcome since it implies that SFLG is not being used to support inferior quality businesses. To assess wider contribution of SFLG, the SFLG group is compared to two groups; conventional borrowers and non-borrowers. The latter group allows some assessment to be made of the benefits of bank finance overall to businesses looking to grow

The CBA is carried out using HMT Best Practice as highlighted in the Green Book. The Cost-Benefit Analysis was conducted using findings gathered from the evaluation survey as well as from Management Information BERR.

The impact evaluation 'assesses the difference between the sample groups holding all other factors constant using econometric modelling techniques'. No further information provided.

Internal validity

No information provided

Inference

No information provided

External validity

No information provided

Cost effectiveness

Calculates live businesses as gross SFLG loans less number that default. Then adjusts live businesses for proportion of borrowers that would have other sources of funds plus businesses that indicate they are simply displacing other firms (based on self reported additionality and displacement). These calculations suggest 55% of supported live firms are finance additional. Combine this with information on self-reported employment additionality to get figures for extra jobs. Similar process for net additional sales, which are then grossed up in to GVA figures using ABI figures on GVA to sales. Attempts to adjust benefits to exchequer by changes in tax, national insurance and welfare receipts.

Overall assessment

This is a difficult report to assess because there is very little technical detail provided. In principle, matching on the basis of a limited number of firm characteristics may help partially correct for selection in to the scheme. But these firms are likely to differ on other unobservable characteristics (reflected in the fact that they were unable to get loan financing under commercial terms). Providing a more detailed assessment is not possible in terms of the detail provided. If the group could be demonstrated as comparable this approach could be rated 3 on the Maryland scale. However, as it stands, this is not demonstrated and, combined with the lack of technical detail, the report would rate 2 on the Maryland scale.

Turning to the cost-effectiveness calculation this uses self-reported estimates of the additionality of the finance, coupled with self-reported estimates of the additionality of the jobs created. Aside from the general problem with self-reported estimates of additionality, it is not clear why you would want to do the analysis in two steps. If SFLG firms can be successfully matched to non-SFLG firms in a way that helped address selection then the most straightforward impact assessment should be based on observed employment differences between the two sets of firms (possibly incorporating additional information on additionality of the loans themselves). It is not clear from the report whether this was considered and if not, why not.

International comparators

Documents examined

Economic Evaluation of The Small firms Loan Guarantee (SFLG) Scheme
www.bis.gov.uk/files/file54112.doc

Appendix: Evaluations of Spatial Policy

This appendix provides details of the evaluations considered in the area of spatial policy. The structure of the template was agreed following discussions with the National Audit Office. In completing the templates, for reasons of both feasibility and presentation, we have made use of source material from the original evaluations without any attempt to provide detailed attribution (e.g. through the use of quotes, or the provision of page numbers).

Local Enterprise Growth Initiative

Policy objectives

To 'release the economic and productivity potential of the most deprived local areas across the country through enterprise and investment – thereby boosting local incomes and employment opportunities and building sustainable communities'.

Scope of evaluation

To describe the activities and outputs attributable to LEGI; to measure and assess the outcomes and impacts of LEGI; to assess the strategic and operational fit of LEGI within the wider policy environment; to identify and share innovation and good practice.

Overall methodology

Profiling of LEGI areas using a range of indicators; Econometric modelling using a difference-in-difference framework to examine changes at neighbourhood level in LEGI and non-LEGI areas in terms of worklessness and business formation; Analysis of programme management information from 20 partnership areas (including interviews with each area's programme manager); Review of existing local evaluation and other research material; Interviews with regional and national stakeholders; Intensive research in six case study areas; Survey of over 560 beneficiary businesses; An assessment of value for money using both 'top-down' and 'bottom-up' data and analyses

Impact evaluation

An overview of the change in business formation and worklessness rates in LEGI areas relative to the national average; a top-down estimate of impact using econometric modelling to identify whether LEGI has had a statistically significant impact in terms of key indicator change in the programme areas; bottom-up estimate of net additional impact using gross performance management data qualified by use of bottom-up evidence [based on interviews and surveys as detailed above].

Policy details

Competitive bidding by Local Authorities (singly or in partnership); two rounds of awards (February 2006, December 2006); flexible activities (LAs choose consistent with objectives); multiple delivery partners. Eligibility depends on ranking on various Indices of Multiple Deprivation.

Data

Firm data from the BETA model (an extensive longitudinal business database, underpinned with data collected since April 1999 to April 2010 from 2.6 million establishments listed with Yellow Pages). Worklessness data from Department for Work and Pensions Longitudinal Database

Costs

Database of activities, spend and outputs assembled from individual area's performance management data used to generate a common programme wide typology of activity. Results collated from the quarterly performance reports from the LEGI areas and verified by project managers.

Outcome variables

Outcome variable defined as annual average growth in worklessness and business formation. The main report suggests time period for worklessness is 2000-2009, gross business formation for 2003-2009. It appears that the report time averages annual growth 200x-2006 as pre-treatment and 2007-2009 as post policy (although this is not always clear).

Control group

Lower Level Super Output Area (LSOA) level analysis. Control group defined using propensity score matching on 'the basis of a range of data including worklessness, population churn, ethnicity, tenure, skills, house prices, crime and working age population.' Unclear if this is the full list of covariates (p.9 of the appendix contains a longer list, but still suggests these are 'examples' of the variables used; text on that page suggests shorter list used for propensity score matching, longer list used as set of controls). Not clear if one-to-one matching. Formula suggests NOT one-for-one, but indexing difficult to interpret.

Methodology details

Difference-in-difference LEGI versus control group. Policy on period is post-2006. Report talks about 'matching' at LSAO level of treatment and controls (see above). No detailed results presented. No summary statistics (or info on number of observations; treatment of errors etc)

Internal validity

No random assignment. Propensity score matching based on observables (although very few details provided). No further discussion of selection bias; No discussion of history; Treatment attrition not an issue (all LEGI LAs spent money). Measurement attrition not an issue (data

for all LSOA and LA). Possibility of maturation (successful LA's might be those that also implement other policies). No specific problems with respect to timing, outliers or repeat testing.

Inference

Basic results are reported with standard errors, but there is no further discussion of inference issues.

External validity

There is very little informal discussion (and no formal discussion) of external validity. Displacement and multipliers dealt with through beneficiary surveys.

Cost effectiveness

Diff-in-diff suggests no impact on worklessness, but positive impact on business formation. Diff-in-diff estimates used to get net additional business from gross business formation in treated areas. The average employment and Gross Value Added per business formed derived from the beneficiary survey are then used to estimate the employment and Gross Value Added impact. The Gross Value Added figure adjusted to allow for 'capital consumption' in order to provide an estimate of Net Value Added. This is then adjusted to allow for persistence (using LEGI area average business survival rates per annum), displacement, multipliers and deadweight (all from beneficiary surveys) to estimate the net additional employment, Gross Value Added and Net Value Added created.

Overall assessment

For a variety of reasons discussed at length in the report (e.g. local flexibility, timing, data) LEGI was always likely to be a difficult policy to evaluate. It could be argued that these problems have been compounded by the fact that the evaluation had multiple objectives: establishing the pattern of spend; undertaking an impact and cost-effectiveness evaluation; considering governance and management arrangements; making policy recommendations. In theory, there may be synergies between these different components, but in practice it is not clear that these are in evidence in the final report.

Turning specifically to the impact and cost-effectiveness parts of the study, the report chooses to adopt a reasonably robust approach – at least in comparison with other UK government evaluations of spatial policies (as detailed elsewhere in this appendix). Specifically, it identifies a control group of LSOA (small areas) based on propensity score matching on observables and uses these in a difference-in-difference analysis. Outcome variables are worklessness and business formation. As discussed in more detail above (see sections on internal validity, inference, external validity) the major problem lies in the implementation and write up of results. In short, the report provides far too little information for an informed reader to establish whether or not the results are valid on any of the main criteria that one

would use to assess the quality of the impact evaluation. The overall approach would rate 4 on the Maryland scale (although in practice, so little information is provided on the matching procedure that one cannot be sure the control and treatment groups are comparable which would imply a ranking of level 2-3 on the Maryland scale).

These estimates underpin the cost-effectiveness evaluation because they are one of the components used to go from gross to net business formation, so problems with the impact evaluation are transferred across to robustness of the cost-effectiveness assessment. The other components used in the cost-effectiveness calculation come from the beneficiary survey. These include figures on average employment and value added and beneficiary 'guesstimates' of the extent of displacement, multiplier and deadweight. It would have been nice to see some attempt to benchmark the average employment and value added figures against other more representative sample data. Estimates of displacement and multiplier could have been derived from the difference-in-difference estimations by widening the area over which LEGI is assumed to have an effect (i.e. by re-running the analysis assuming that 'nearby' LSOA are also treated by LEGI). It is worth noting, that this observation also raises concerns about whether or not the propensity score matched LSOA include those in nearby LSOA (these would underestimate the treatment effect in the presence of multipliers, overestimate in the case of displacement). Finally, timing of effects would seem to be an issue that receives very little consideration in the report.

In terms of improving the evaluation, the first order issue would have been the provision of much more information to allow a fairer assessment of internal and external validity. Taking a broader perspective, the estimates of additionality could have been further refined by using the 'thresholds' incorporated in to the policy (e.g. areas just invalid for the policy on the basis of IMD ranking, or because of location outside a LEGI boundary could be used as a good control group). Given the difficulties in assessing the quality of the impact evaluation and cost-effectiveness based on the information provided it would be difficult to be confident in making changes to the policy on the basis of the evaluation. In practice, this is a moot point because the coalition government stopped the LEGI scheme around the same time as this evaluation was published. To the best of our knowledge, this evaluation has had little impact on the proposals for Enterprise Zones (which might be seen as the closest successor policy).

International comparators

US Enterprise/Empowerment Zones.

Documents examined

LEGI National Baseline
http://www.communities.gov.uk/documents/communities/pdf/1098905.pdf LEGI update report http://www.communities.gov.uk/documents/communities/pdf/1097253.pdf LEGI National Evaluation final report (plus appendices)
http://www.communities.gov.uk/documents/regeneration/pdf/1794470.pdf
http://www.communities.gov.uk/documents/regeneration/pdf/17994501.pdf

The Mixed Communities Initiative

Policy objectives

The Mixed Communities Initiative was seen as a new approach to tackling area deprivation in England. Its distinctive characteristics were: aimed at fundamental long term transformation rather than more modest improvements; it emphasised changes in population mix; and it was dependent on local private/public partnership rather than on a 'cash pot' from central government.

Scope of evaluation

12 local 'demonstration projects. Five key issues: objectives and how they vary in different areas; whether the approach is deliverable; whether it is an effective way of delivering new affordable housing and Decent Homes; what initiative adds as a new model of regeneration; and how benefits to existing residents can be secured at least cost. Specific issues: to clarify overall objectives of the Mixed Communities Initiative; to identify a set of common measures against which demonstration projects can measure and assess their progress; to establish whether demonstration projects have been successful in meeting their aims, in the period 2006-2009, and the reasons for success and barriers to success; to identify transferable lessons.

Overall methodology

The methodology for the evaluation included stakeholder interviews to establish MCI objectives; case studies of six demonstration projects; lighter touch monitoring of progress in the other demonstration projects, and analysis of quantitative data on area change.

Impact evaluation

Comparisons of key indicators to national trends, local authority districts and comparator areas within local authority districts

Policy details

Government designated 12 existing or planned local schemes as demonstration projects. Chosen following recommendations from Government Offices for the Regions, based on the criteria that demonstration projects should have clusters of super output areas (SOAs) in the 2 per cent most deprived neighbourhoods in England, and that if possible there should be one from each region. The demonstration projects covered diverse areas with projects at different stages.

Data

ONS experimental statistics on small area population; Number of dwellings based on council tax records.

Outcome variables

Relevant output measures include numbers of homes: in different tenures; in different price ranges; failing Decent Homes standards. Outcome measures include: house sales volumes and prices; lettings periods for social housing; numbers and characteristics of in-movers and out-movers are also important indicators. Core basket of indicators, covering education, employment and health, including: educational attainment at Key Stage 2 and GCSE; Jobseeker's Allowance and Incapacity Benefit counts and flows; rates of Coronary Heart Disease, infant mortality and mental ill-health.

Control group

Local authority districts in which they are situated and 'comparable areas' (all the similarly disadvantaged areas in the top 5% of the national Index of Multiple Deprivation) within those local authority districts. These figures include the demonstration project areas.

Methodology details

Figures showing trends relative to comparators. Very preliminary (in terms of timing) but no basic regressions or further econometric analysis.

Internal validity

No impacts expected at stage of evaluation. Report suggests that it can say little about actual outcomes at this stage, since changes in outcomes would not be expected to have been achieved by this stage of the developments. Awareness of displacement, but not formally considered.

Inference

Not applicable.

External validity

Not applicable.

Cost effectiveness

Not applicable.

Overall assessment

This is an interesting case study in the context of the current project because the pilots were set up using criteria that would allow in-depth assessment of *process* rather than impacts. As such it does not have an obvious ranking on the Maryland scale. Even if 'evaluation' had been done later, it is very hard to see how this could have been used to assess impact given way pilot areas were chosen rather than randomised. This suggests that a careful analysis of the pilot would likely to badly on the Maryland scale because of the lack of an appropriate comparison group. This provides an interesting contrast with other policy areas (e.g. labour market) where pilots were seen as a way of helping establish that the policy had the desired impacts. This focus on process over impact is a consistent theme arising from the evaluations reviewed that cover spatial policies. It would be useful to understand why these 'cultural' differences arise between sets of evaluations covering different policy areas.

International comparators

Hope IV US

Documents examined

Evaluation of the Mixed Communities Initiative: Demonstration Projects - Final report
http://www.communities.gov.uk/documents/housing/pdf/1775216.pdf

National Strategy for Neighbourhood Renewal: Final report

Evaluation of the National Strategy for Neighbourhood Renewal: Econometric modelling of neighbourhood change

Policy objectives

The National Strategy for Neighbourhood Renewal (NSNR) launched in 2001 with the vision that: “within 10 to 20 years no-one should be seriously disadvantaged by where they live”. It had two long-term goals: “in all the poorest neighbourhoods to have common goals of lower worklessness and crime, and better health, skills, housing and physical environment” and “to narrow the gap on these measures between the most deprived neighbourhoods and the rest of the country”.

Scope of evaluation

Final report examines the extent and nature of neighbourhood deprivation with particular reference to 2001; examines how conditions have changed since that time and the factors that appear to have been particularly significant in influencing that change; assesses the degree to which that change appears to have been attributable to NSNR and the extent to which the Strategy has represented value for money; examines the effectiveness and relevance of the different structures and tools introduced or adopted by the Strategy; summarises lessons learned. **Econometric modelling of neighbourhood change** describes econometric 'transition model' developed for worklessness that forms the basis of top-down VfM calculations for worklessness. A separate appendix (see below) considers the evaluation on the NRF impact on educational outcomes.

Overall methodology

Final report: Develops a typology of neighbourhood types: isolate, transit, escalator, gentrifier; provides descriptive statistics on absolute and relative gaps relative to benchmarks; summarises results from two underlying evaluations on neighbourhood change (which form the main focus of this appendix) and educational outcomes (considered in a separate appendix below). Reports results from 'Top-down' (i.e. econometric model based) and 'bottom-up' (i.e. based on 'informed' assessment of additionality) estimates of VfM. Also considers governance processes.

Impact evaluation

Neighbourhood change: Logit modelling of transition matrices based on discretised relative worklessness rates for LSOAs.

Policy details

IMD used to select local authority areas for receipt of NRF (and subsequently WNF). LAs free to develop own policy priorities (elements of local consultation).

Data

Final report: Descriptive statistics from variety of data sources. **Neighbourhood change model:** lower layer super output area (LSOA) data supplied by the Social Disadvantage Research Centre (SDRC) at Oxford University.

Costs

Constructed as part of the evaluation. There are few figures available for spend of NRF between the various domains. The Fund allowed flexibility for decision-making at a local level as to the neighbourhoods and the interventions that should receive funding. It was intended, in effect, as a top-up to local areas, to help them to begin improving core services in their most deprived neighbourhoods, rather than as a conventional 'programme'. It was not ring-fenced and reporting arrangements – and hence any central collation of management information – were limited.

Outcome variables

Neighbourhood change model: Worklessness rates banded in to 20 groups and used to estimate transitions.

Control group

Final report: Descriptive statistics on (1) difference between NRF local authority districts (LADs) and the national average; (2) difference between the most deprived LSOAs and the rest nationally (3) difference between the most deprived LSOAs and the rest within local authority areas. **Neighbourhood change model:** Not explicitly defined but appears to be LSOA with similar ranking relative to own LA and similar observable characteristics but who do not receive NRF or NDC money.

Methodology details

Neighbourhood change model Defines appropriate transition matrices for worklessness using LSOA data; construct and estimates binary models of transition employing discrete dependent variable based methods. Uses estimates to examine role of underlying social, economic and policy factors in explaining the transition process. Area transitions are defined for neighbourhood worklessness rates relative to the average Local Authority District (LAD) rate at different points in time. For base year (2001), all 32,482 English LSOAs are ranked according to the ratio of their worklessness rates with host LAD values and grouped into 20 bandings each of which corresponds to 5 percentile points on the overall distribution. The same approach is adopted for 2006 with the banding thresholds used to segment the distribution remain fixed at the 2001 ratio values. This approach constructs a 20 by 20 transition (or origin/destination) matrix showing the extent to which LSOAs remain in their original band or move to higher or lower bands between 2001 and 2006. Binary dependent variables are then constructed depending on LSOA transitions upwards or downwards across bands. Modelling concentrates on two types of transition: (1) probability of improvement among LSOAs defined to be within the bottom 4 bands (i.e. 20%) in 2001; (2) probability of a deterioration into the bottom 5 per cent or 10 per cent of LSOAs in 2006 from 'better' band positions in 2001. Logit modelling on transitions as a function of a large number of control

variables (including NDC and NRF policy dummies; although the latter are hard to define given lack of spatial targeting). Report provies extensive discussion of measures of goodness of fit. VfM model replaces binary indicators for policy with some measure of amount of expenditure by different policies (although see below on problems with cost data).

Internal validity

Neighbourhood change model: No random assignment. Logit transition model controls for observable characteristics. No discussion on selection bias (although final report confirms that areas had different pre-trends and not clear extent to which this is captured in analysis). No discussion of history or timing; Treatment attrition not an issue (all NRF LAs spent money). Measurement attrition not an issue. Possibility of maturation (successful LA's might be those that also implement other policies). No specific problems with respect to timing, outliers or repeat testing. Not much other information provided.

Inference

Neighbourhood change model: Basic results are reported with standard errors, but no further discussion of inference issues (in contrast, there is extensive discussion of goodness of fit criteria)

External validity

Neighbourhood change model: Some robustness checking on outcome variable: 12 different models estimated depending on exact definition of positive or negative transition (i.e. between which bands). Different estimations undertaken using dummy for treatment versus estimated expenditures. There is very little informal consideration (and no formal discussion) of external validity. Displacement and multipliers not addressed.

Cost effectiveness

Neighbourhood change model: Uses differences in the probability of an area transiting from one band to another to derive estimates of what worklessness levels in NSNR areas would have been in the absence of the policy, all other characteristics of the area being unchanged. For example, if an NRF area was 20 per cent more likely to improve and the actual number of individuals who were no longer workless was 50, then we can say that the policy effect had led to 10 fewer people being workless. Bottom up: The analysis was based largely on the judgement of the evaluation team, drawing on information from project and programme managers, and supported, where possible, with the views of NRF coordinators as well as beneficiary and other information (where available). Given almost *no* idea of expenditures (see above), so assumptions heroic.

Overall assessment

For a variety of reasons discussed at length in the report (e.g. local flexibility, timing, data) NRF was always likely to be a difficult policy to evaluate. As with the evaluation of LEGI, it could be argued that these problems have been compounded by the fact that the evaluation had multiple objectives: establishing the pattern of spend; undertaking an impact and cost-

effectiveness evaluation; considering governance and management arrangements; making policy recommendations. In theory, there may be synergies between these different components, but in practice it is not clear that these are in evidence in the final report.

It should be highlighted that the lack of effective programme management data on expenditures presents a very severe problem for the evaluation of NRF. The difficulties of capturing this data are clearly conveyed in the final report.

Turning specifically to the impact and cost-effectiveness parts of the study, the neighbourhood change model is *very* difficult to interpret. Control groups are not carefully identified and there is very little information provided on which to assess the methodology and results. As a result, at best, the approach ranks 2 on the Maryland scale.

In addition to problems arising from this lack of detail, one has to question whether the overall approach taken is useful for establishing cost-effectiveness. The methodology (described above) treats the estimates of the extent to which treatment affects propensity to transit groups as an estimator of additionality. This is very difficult to understand and it is not clear that this is in any way meaningful. These estimates underpin the cost-effectiveness evaluation because they are the central component used to go from observed worklessness changes to net changes attributable to the programme.

The cost-effectiveness calculation ignores the extent of displacement, multiplier and deadweight. As with the LEGI evaluation estimates of displacement and multiplier could have been derived from the difference-in-difference estimations by widening the area over which NRF is assumed to have an effect. Finally, timing of effects would seem to be an issue that receives very little consideration in the report.

In terms of improving the evaluation, the first order issue would have been to properly justify the use of the transition model. Our assessment is that it would have been much more transparent to estimate a model for levels of worklessness (with LA levels as one control variable) and to use this as the basis for the evaluation. Taking a broader perspective, as with LEGI, the estimates of additionality could have been further refined by using the 'thresholds' incorporated in to the policy (e.g. areas just invalid for the policy on the basis of IMD ranking could be used as a good control group). Given the difficulties in understanding the modelling approach adopted, it would be difficult to be confident in making changes to the policy on the basis of the impact and cost-effectiveness provided.

International comparators

Documents examined

CLG (2010) Evaluation of the National Strategy for Neighbourhood Renewal: Local research project; CLG (2010) Evaluation of the National Strategy for Neighbourhood Renewal: Modelling neighbourhood change;
http://www.communities.gov.uk/publications/communities/evaluationnationalchange

Evaluation of the National Strategy for Neighbourhood Renewal: Improving educational attainment in deprived areas.

Policy objectives

Improving the skills of people living in the most deprived neighbourhoods is one of the five priority goads of the National Strategy for Neighbourhood Renewal (NSNR). It has been estimated that about 20 per cent of the spending on the Neigbourhood Renewal Fund has been on education interventions.

Scope of evaluation

This report describes the econometric 'difference-in-difference model' developed for educational outcomes. Outcomes are measured between 2002 and 2006.

Overall methodology

Sections of report provide: summary of aggregate level attainment rates for children in different groups to show the extent of the gap that exists between children in deprived local authority areas and comparator benchmarks; explanation of data and methods used – including how treatment groups and control groups selected (through statistical matching); shows main results and analysis by sub-group (gender, ethnicity, region); summary of key results and main messages.

Impact evaluation

Difference-in-difference models applied on schools thought likely to be the target of treatment and selected control schools in non-NSNR areas. Controls are included in the regression for observable characteristics.

Policy details

Although education interventions accounted for much of NSNR spending, the evaluators do not have much information about this. They say 'one issue with which the evaluation had to contend is that there is no clearly defined treatment group identified within the policy to which interventions should be directed, nor in practice is there knowledge about how interventions in NSNR districts have been targeted'. This is why they create four possible treatment groups where one might expect interventions to have been directed. They choose schools as the appropriate unit because they are commonly used to target education interventions.

Data (appropriate, collected, used)

Administrative pupil-level data from the National Pupil Database (2002-2006) linked to the school and LA Information System (LEASIS).

Costs

Not discussed in this report.

Outcome variables

Nine different outcome measures consisting of four outcome measures at Key Stage 3 (age 14) and five outcome measures at Key Stage 4 (age 16).

Control group

Schools in NSNR areas selected to look as similar as possible to schools used for treatment. This is based on propensity score matching.

Methodology details

Uses administrative data on pupils in secondary schools from 2002 to 2006 [National Pupil Database]. Define four possible treatment groups in NSNR areas, deemed likely to have received some support from NSNR. These are all defined according to school characteristics rather than pupil characteristics. The researchers think it is likely that interventions would have occurred more at school level than targeted at specific pupils across different school types. The four treatment groups are overlapping (about 70% of schools within any one treatment group will also be in another treatment group). Group 1 is pupils within the 25% most poorly performing schools in each NSNR district according to each school's Key Stage 3 score attainment in 2002. Group 2 is similar but defines 'poorly performing' on the basis of Key Stage 4 (GCSE) results in 2002. Groups 3 and 4 identify treatment groups based on the most highly disadvantaged schools (index of multiple deprivation).

The control schools are selected from neighbourhoods not exposed to the NSNR using propensity score matching. However, unlike the programme evaluation literature which uses this method to select treatment and control schools that have 'common support', this method is used to trim the sample to omit potential 'control schools' that look much too different. The sample of treatment schools is not trimmed and most 'treatment' schools look as though they do not have a valid comparator in control schools based on observable characteristics.

A 'difference-in-difference' estimation is then conducted on the selected schools where 2002 is treated as the 'pre-policy year' (although the NSNR was introduced in 2001). The outcome variables related to educational attainment at the end of Key Stage 3 (when students are about 14 years of age) and the end of Key Stage 4 (GCSEs or equivalent).

Internal validity

There are several problems with this:

1) Not certain that interventions were targeted at school-level and at these schools.
2) Propensity score matching not used to restrict sample to 'common support' but used to trim the sample of control schools. Treatment schools still look more disadvantaged on observable methods.
3) The year selected for 'pre-treatment' was in fact the second year of the NSNR policy. This should bias results downward.
4) Schools selected as treatment and controls might be trending differently in pre-policy period. It is possible that positive results attributed to the policy are actually attributable to (a) regression to the mean; (b) other school/regional/national policies targeted at the poorest or lowest performing schools over this period.

Inference

Most details given in the appendix.

External validity

Same problems as discussed above. No explicit discussion.

Cost effectiveness

Costs are not discussed. Positive impacts are found both at Key Stage 3 and Key Stage 4. The results are estimated to represent an average improvement of about one-tenth of one level in each subject at Key Stage 3. At Key Stage 4, consistent and significant improvements in attainment were also found. In some cases, the positive impacts are apparent in later years and not earlier years. This is interpreted to mean that the positive impact increases over time.

Overall assessment

Overall, this report is a considerable improvement over the other spatial policy evaluations that we have considered. The approach adopted would rank 4 on the Maryland scale if carefully implemented, although problems in implementation mean that in practice it would be more appropriate to rank this as level 2. It is interesting to note that overall the report is notably weaker than the majority of the education evaluations that we have considered, even if the approach adopted is reasonably robust.

NSNR is a difficult policy to evaluate in terms of education impact because there are no clear details on what exactly happened as a result of the NRF expenditure. The researchers are targeting schools that they think are likely to have been the subject of intervention. However, more of the spending could have been at primary level (not considered here) or targeted as families/neighbourhoods rather than at schools. Many children from disadvantaged families will attend schools that are not classified as most deprived or lowest performing.

The analysis is unusual for not using propensity score matching to target schools that have 'common support' and instead uses it to trim the sample of control schools. This matters because the most deprived (treatment) schools do not have counterparts and it is very possible that the positive results attributed to NSNR are in fact attributable to (1) other policies targeted at the poorest and/or lowest performing schools; or (2) regression to the mean.

The analysts could have looked as pre-programme trends in outcomes for treatment and control schools (they have data on pupil attainment that precedes 2002) but they did not do this.

Using 2002 as a 'pre-intervention' year is strange since this is actually one year post-intervention. The researchers did this because some pupil-level information (e.g. ethnicity, free school meal status etc.) was only collected from 2002 onwards. However, it was possible

to get pupil level attainment data (matched to their previous attainment) before this time. So the sensitivity of results could have been checked.

International comparators

Not clear.

Documents examined

CLG (2010) Evaluation of the National Strategy for Neighbourhood Renewal: Improving educational attainment in deprived areas.
http://www.communities.gov.uk/documents/communities/pdf/1490497.pdf

Regenerating the English Coalfields – interim evaluation of the coalfield regeneration programmes

Policy objectives

Covers three regeneration programmes: The National Coalfields Programme (NCP), the Coalfields Regeneration Trust (CRT) and the Coalfields Enterprise Fund (CEF). These schemes aimed to lay the foundations for sustainable regeneration of the former coalfield areas through physical reclamation and renewal, community capacity rebuilding and human capital development, and the promotion of enterprise and business growth.

Scope of evaluation

Multiple: overall progress of coalfield areas; the range of problems being addressed by the specific coalfield programmes; integration of coalfield problems in regional policy; implementation of the coalfield programmes; impact of the DTI coal health compensation scheme; additionality, displacement and other adjustments to programme outputs; cost-effectiveness and value for money; impact of the programmes; inter-generational outcomes in coalfield areas.

Overall methodology

There were four broad strands to the evaluation: a review of the literature; an analysis of secondary data sources since 1998; an assessment of regeneration programme documentation and monitoring data; and six case studies reviewing the changing conditions and the influence of the programmes in the local areas. Consultations with 134 regeneration partners; 36 project managers; 28 property developers; survey 1332 households; survey 602 businesses.

Impact evaluation

Additionality/displacement drawn from interviews with programme managers, partners and beneficiaries as revealed from consultations, surveys and reviews of relevant evaluations.

Policy details

Three strands of funding administered by the Department for Communities and Local Government (DCLG), namely the Homes and Community Agency's (HCA) National Coalfields Programme, the Coalfields Regeneration Trust (CRT) and the Coalfields Enterprise Fund (CEF).

Data

The secondary data for each of these three spatial levels were constructed by building up from the finest grained spatial level – coalfield wards or SOAs. Any variables that could not be derived in this way were not included in the data set. This was because spatial levels wider than wards or SOAs (such as Local Authority Districts) would embrace areas other than coalfields and, therefore, data at this level could be misleading about the conditions prevailing in the coalfields. Various sources: ABI, LFS, DWP benefits, census 2001, IMD, ONS Neighbourhood Statistics.

Costs

The monitoring system used by EP (for NCP) monitors both financial and output data on a project by project basis. As projects are site-specific, the monitoring system also attributes spend and outputs to the ward in which the site is located. Hence, it is possible to evaluate both spend and outputs of the NCP at a programme level and also at the scale of individual coalfield sub-regions with sites allocated to the latter. The Trust uses a monitoring system that records all spend and output data against the individual project. The CRT was not required to attribute spend and output data to geographically defined areas such as local authority districts or wards, although projects were required to state the geographical footprint in which the project would deliver (district level in Round 1, and ward level in Rounds 2 and 3). The system, therefore, is not able to attribute the level of spend or division of outputs between coalfield areas. For the purposes of the evaluation, the CRT divided the total spend on a project by the number of districts to which the project delivered, to produce broad-brush estimates of spend by district. Enterprise Ventures (for CEF) submit a quarterly financial monitoring and progress report to CLG. At the time of the report a total investment of £1.76m has have been made in nine businesses. The programmes' expenditure and outputs were likely to be exceeded by non-coalfield specific regeneration programmes. The expenditure of the programmes is also compared where possible with the expenditure of the non-coalfield specific regeneration programmes. It was not possible to compare the coalfield regeneration programme expenditures with spend on non-coalfield specific regeneration programmes (such as the Single Regeneration Budget (SRB) and the New Deal for Communities (NDC)) other than at the level of Local Authority Districts.

Outcome variables

Comparisons of levels and changes (where available) for employment and other labour market indicators.

Control group

Three coalfield spatial levels are used in the assessment – all coalfield areas in England, sub-regional coalfields and case study local coalfields – contrasted with the average for non-coalfield England and the regions. Performance of the coalfield sub-regions (relative to all English coalfields) against the performance of the non-coalfield regions in which they are located (relative to non-coalfield England as a whole) used to assess to what extent the

pattern of adjustment and conditions in the coalfields a function of regional differences across England.

Methodology details

Adjustment from gross to net outputs based on review of programme documentation (e.g. the evaluation of the CRT) and from consultations with project managers and property developers/agents and the survey of businesses carried out for the six case studies. All of these figures are self-reported. The report uses an overall additionality rate of 75 per cent (the mid-point of the 70-80 per cent range reported) to derive the net outputs of the NCP and CRT. This is the local additionality rate – not a sub-regional or regional rate, which the reports suggest would be expected to be much lower. [The definition of coalfields in the analyses below follows the methodology set out by Sheffield Hallam in 2004, based on 2003 ward boundaries, and converted to Lower Super Output Area (LSOA) boundaries.]

Internal validity

The robustness of the additionality rates assessed for the programmes was checked against the accumulation of evaluation evidence for similar programmes (such as the Single Regeneration Budget). Displacement from other areas not considered (see comment above about use of local multiplier).

Inference

Not applicable.

External validity

In some senses, not applicable – programmes cover all coalfields. Would be more of a concern if results were to be used to inform development of wider regeneration policy (given that external validity not considered).

Cost effectiveness

Comparison of net cost per job estimates to English Partnership benchmarks. Comparison of actual job changes to job changes possibly attributable to expenditure. The assessment was constrained by secondary data limitations and because programme outputs were not necessarily specified that could easily be related to changes in the relevant coalfield condition indicators

Overall assessment

It should be noted at the outset that these are a difficult set of policies to evaluate. There are multiple funding streams, with multiple objectives and all of the relevant coalfields are treated. That said, there were several rounds of expenditure and the intensity of treatment varied across rounds. This variation could have been utilised to get much better estimates of the impact of the policy on outcomes of interest. It might also have been appropriate to benchmark against comparable non-coalfield areas using a variety of socio-economic indicators. The reliance on self-reported additionality is problematic for many of the standard reasons (discussed below in the context of the evaluation of Regional Development Agencies). As additionality is central to the impact evaluation, the use of self-reported figures means that overall this report would rank at level 1 on the Maryland scale.

These problems of using self-reported additionality are compounded in this situation where there are multiple objectives. It's also not clear that self-reported additionality is used consistently in the net cost per job estimates. For example, the report appears to apply building net additionality (of 75%) to *jobs* for people located in those buildings (even though that is inconsistent with survey of business). More could have been done to consider whether

this was appropriate. The resulting additionality figures for employment also appear large for comparable self-reported numbers for employment impacts of similar types of expenditure for RDAs. When turning to the cost-effectiveness calculations these problems are compounded, by the fact that for CRT no monitoring of expenditure or outputs by area. This appears to be a recurrent them for area based programmes (see, for example, appendix on NRF) and makes any evaluation very difficult.

International comparators

Documents examined

Coalfield regeneration review board (2010): A review of Coalfields regeneration
http://www.communities.gov.uk/documents/regeneration/pdf/1728082.pdf

Supporting documents consulted where necessary (e.g. for clarification):

Exosgen (2010) Evaluation of the Family Employment Initiative
http://www.coalfields-regen.org.uk/docs/199.pdf

A mine of opportunity: local authorities and the regeneration of the English coalfields (2008) Audit Commission; Regenerating the English Coalfields (2009) National Audit Office

Impact of Regional Development Agencies spending

Policy objectives

RDAs aimed to further economic development and regeneration; promote business efficiency, investment and competitiveness; promote employment; enhance development and application of skills; contribute to sustainable development

Scope of evaluation

To understand purpose of RDA interventions; map RDA spending on each intervention, identify gross outputs and assess the extent to which these were additional; determine outcomes and impacts associated with the net outputs; assess value for money.

Overall methodology

Methodology involved a review of over 640 individual RDA evaluations with work done to standardise GVA impacts (actual annual and cumulative and future potential) for over 400 evaluations undertaken before national guidelines issued.

Impact evaluation

The national report and regional annexes provide no overview of the methods used to move from gross outputs to net outcomes. A partial review of some of the underlying evaluations suggests that two approaches have been adopted. Some of the underlying evaluations ask businesses or programme managers about additionality and use these figures to move from gross to net outcomes. Other reports do not directly address additionality, instead using

figures either from other evaluations from the same RDA or from previous national evaluations (particularly those from the Single Regeneration Budget evaluation).

Policy details

RDAs undertook a wide range of interventions that the national evaluation classifies in to business, place, people or other. See the report for details.

Data

Many sources depending on underlying evaluation.

Costs

Level of detail in the report suggests that there is relatively good data on costs (although the extent to which this spend is covered by suitable evaluations varies by RDA and by theme).

Outcome variables

Uses wide variety of outcome measures depending on area of policy spend and intended outputs. These are then translated in to GVA figures (see above)

Control group

Almost impossible to assess because report provides no overview of methodologies adopted in the underlying evaluations. It would appear that most of the underlying evaluations do not use any control group and instead rely on self-reported evaluations (either generated by the evaluation or taken from other reports – see above).

Methodology details

See above – most underlying reports appear to rely on self-assessed additionality (with all the problems that this entails)

Internal validity

Not applicable

Inference

Not applicable

External validity

Not applicable

Cost effectiveness

The value for money calculation involves moving from outputs to net employment and then uses net-employment combined with estimates of GVA per worker to calculate actual (annual and accumulated) and potential future GVA contributions. This seems a reasonable approach for coming up with some way of comparing across disparate themes although (i) it is biased against interventions that create non-employment impacts and (ii) it is reliant on the validity of the methods used to move from gross outputs to net outcomes.

Overall assessment

This is a very frustrating report. There is a vast amount of detail (running to over 500 pages) but the report makes no attempt to systematically describe the methods used in the underlying evaluation reports. As such, it is very difficult to assess the quality of the evidence on the basis of the report outline. Further investigation of the underlying evaluations suggests a heavy reliance on self-reported additionality. These additionality figures may be generated by the evaluation itself or taken from other evaluations. This approach would rank as level 1 on the Maryland scale.

In terms of the consistency across evaluations, there is a surprising degree of variation in additionality across projects given the heavy reliance on 'benchmark' figures taken from, for example, the Single Regeneration Budget evaluation. The more general problem, relates to the systematic use of self-reported additionality provided either by recipients of the money or by people directly involved in handing out the money. Most academic experts on policy evaluation would view this figures as being highly unreliable. They require firms of programme managers to be able to accurately evaluate the counterfactual - i.e. what would have happened in the absence of the intervention. This is a very difficult thought experiment at the best of times and one that is made more difficult with policy evaluation because the people being asked are often receiving money (or some kind of benefit in kind) from the operation of the policy.

Further cause for significant concern arises when these self-reported additionality figures are used as the basis for comparisons across different policy areas or different types of recipients. Why should we expect a young unemployed worker assessing the additionality of a training scheme to give us numbers that can meaningfully be compared to those from a scheme supporting R&D? More subtly, even within schemes, why should we expect the answers to such questions to be the same across, say, small and large firms? One reason why the answers might differ is because the policy actually differs in terms of additionality for the different types of interventions etc. But more worrying is that the answers might differ depending on characteristics of the policy that have nothing to do with whether the policy has any impact on behaviour.

Unfortunately, we know very little about the direction of biases in practice. These issues raise significant concerns for this evaluation that do not appear to be considered in the report.

International comparators

Various EU regional interventions (Cohesion and Structural Funds)

Documents examined

BERR (2009) Impact of RDA spending – National Report – Volume 1 – Main report

BERR (2009) Impact of RDA spending – National Report – Volume 2 – Regional Annexes

[plus a number of the underlying evaluation reports]

Single Regeneration Budget

Policy objectives

SRB came in to operation in 1994 to encourage partnership working in local regeneration by acting as a flexible funding supplement to main stream programmes. Multiple objectives included enhancing employment prospects, education and skills; encouraging economic growth; improving housing through physical improvements; tackling crime and enhancing quality of life

Scope of evaluation

To design a methodology to evaluate the process by which economic, social and physical regeneration achieved; to undertake an evaluation of the impact and cost effectiveness of the first and second rounds of spending; to undertake an analysis of unsuccessful bids.

Overall methodology

Multiple components considered the changing policy response; the targeting of need; the role of partnership working; the development of innovative thematic solutions; leverage on to mainstream funds ('bending'); scheme outputs, additionality and value for money; the joining up of regeneration efforts across themes; the extent outcomes sustainable.

Impact evaluation

The impact evaluation first obtained gross output measures for about 60 outcomes for 20 case study areas. Figures are also provided separately for black and ethnic minority communities for around 40 of these outputs. Additionality is then assessed based on interviews with project managers and partners (plus some results from beneficiary surveys). In effect, the 20 case studies generate over 100 self-reported measures of additionality. A similar process is used to assess the additionality of SRB expenditure (that is to say, the extent to which it displaced other local expenditure that would have happened anyhow). For 7 out of the 20 case study areas a social survey is available before and after intervention and this is used to provide further evidence on benefits (across a huge range of expenditures) in those 7 areas.

Policy details

SRB undertook a wide range of expenditures in a variety of different areas over six different rounds of expenditure. See the report for details.

Data

A huge amount of data is collected for the case study areas (including for 100 different output measures)

Costs

The report describes the large amount of data work needed to get distribution of expenditure by Local Authority (going to finer spatial scales considered infeasible in terms of resource costs). For example, for three out of six rounds, the authors needed to construct the data scheme by scheme from information on 100s of schemes provided on the ODPM web site.

Outcome variables

Uses wide variety of outcome measures depending on area of policy spend and intended outputs.

Control group

None. As described above, all additionality is based on self-reported assessments.

Methodology details

See above – self-assessed additionality on 20 case study areas (with all the problems that this entails)

Internal validity

Not applicable

Inference

Not applicable

External validity

Not applicable

Cost effectiveness

The value for money calculation involves using additionality adjusted gross outputs for selected outcomes relative to the exchequer cost of the expenditure (e.g. cost per job, etc).

Overall assessment

The SRB evaluation was a vast, ten year, undertaking and had many different objectives. For the purposes of the current assessment the crucial issue concerns the robustness of the additionality and value for money calculations. As should be clear from the description

above, the evaluation provides measures of additionality for over 100 outcomes constructed from self-reported assessments for 20 case studies. This approach would rank as level 1 on the Maryland scale. The report does urge caution in the extent to which these figures should be taken as representative, although it does not address the deeper conceptual issues about the appropriateness of self-reported additionality (as described in the appendix for the RDA evaluation). Unfortunately, the note of caution about representativeness does not appear to have stopped the widespread use of these additionality figures in other evaluation reports (again, as described in the appendix for the RDA evaluation).

International comparators

Various EU regional interventions (Cohesion and Structural Funds)

Documents examined

The Single Regeneration Budget – Final Report
http://www.landecon.cam.ac.uk/staff/publications/ptyler/SRB_part1_finaleval_feb07.pdf
http://www.landecon.cam.ac.uk/staff/publications/ptyler/SRB_part2_finaleval_feb07.pdf
http://www.landecon.cam.ac.uk/staff/publications/ptyler/SRB_part3_finaleval_feb07.pdf