ARTICLES

EVALUATION, POLICY LEARNING AND EVIDENCE-BASED POLICY MAKING

IAN SANDERSON

The increasing emphasis on the need for evidence-based policy indicates the continuing influence of the 'modernist' faith in progress informed by reason. Although the rationalist assumptions of evidence-based policy making have been subject to severe challenge from constructivist and post-modernist perspectives, it is argued that the attempt to ground policy making in more reliable knowledge of 'what works' retains its relevance and importance. Indeed, its importance is enhanced by the need for effective governance of complex social systems and it is argued that 'reflexive social learning' informed by policy and programme evaluation constitutes an increasingly important basis for 'interactive governance'. The expanded use of piloting of new policies and programmes by the current UK Government is considered to provide limited scope for evaluation to derive reliable evidence of whether policies work. There is a need for greater clarity about the role of evaluation in situations where piloting essentially constitutes 'prototyping'. More emphasis should be placed on developing a sound evidence base for policy through long-term impact evaluations of policies and programmes. It is argued from a realist position that such evaluation should be theory-based and focused on explaining and understanding how policies achieve their effects using 'multi-method' approaches.

INTRODUCTION

The legacy of the Enlightenment is proving robust against post-modernist attacks on notions of rationality. In spite of the 'rage against reason' (Bernstein 1991) conducted in the name of post-structuralist deconstruction, the forces of optimism about the role of scientific inquiry in pursuit of progress still prevail. The realist tradition of social explanation provides a strong and resilient basis for such optimism in its quest to provide '. . . explanatory purchase on substantive social problems' (Archer 1995,

Ian Sanderson is at the Policy Research Institute, Leeds Metropolitan University.

Public Administration Vol. 80 No. 1, 2002 (1–22)

p. 12). This quest to understand and explain what works for whom in what circumstances underpins the notion of evidence-based policy making in that the latter relies on the assumption that we can make policies work better if we understand how policy mechanisms bring about change in social systems to achieve desired outcomes (Pawson and Tilley 1997).

The optimistic assumptions of modernity, of progress driven by scientific advance, have been defended by Bronk (1998). Bronk argues that we can preserve the force of these assumptions by developing a strong framework of morality, social cohesion and rational, evidence-based government intervention. This modernist-rationalist project is reflected in the drive to reform the public sector across OECD countries over the past two decades. With increasing questioning and scrutiny of public intervention in economic and social spheres, governments are turning to evidence of performance for legitimacy since it is no longer guaranteed solely by democratic political processes. Indeed, the OECD (1994) has argued that 'results-oriented management' provides a new management paradigm.

Concern with the relationship between social science and public policy is, of course, not new. Bulmer (1987) identified the mid-1960s as the take-off point for social science in Britain. The Report of the Heyworth Committee on Social Science recommended an expansion in social science research and led to the establishment of the Social Science Research Council (SSRC). Moreover, the rapid growth in expenditure on public services created a demand for information about the nature of social problems and the effect and impact of public policies (ibid., pp. 3–4). However, the relationship between social science and policy making has always been problematical in the UK due to a political culture which, as Bulmer (ibid., pp. 6–9) argues, has been largely resistant to the influence of 'rational knowledge'. Progress has been made in the health field, on the other hand, since the early 1990s with the establishment of the Cochrane Collaboration, with a centre in the UK, to bring evidence from reviews of the effects of healthcare interventions to bear upon decision making. In 1999 this initiative was emulated in the fields of social and educational policy with the launch of the Campbell Collaboration.

Throughout its life, the SSRC (and subsequently the Economic and Social Research Council – ESRC) has been concerned with the role of social science in informing public policy but has had to acknowledge the continuing struggle to bring the results of social science research to bear effectively on public policy (Newby 1993). In February 2000 the then Secretary of State for Education and Employment, David Blunkett, in a keynote lecture to the ESRC also highlighted a range of problems relating both to the 'relevance' of research and the 'culture of anti-intellectualism' in government (DfEE 2000a). Nevertheless, in this lecture, Mr Blunkett emphasized the commitment of the Labour Government to '. . . using information and knowledge much more effectively and creatively at the heart of policy making and policy delivery . . .' (ibid., p. 2) in the context of the agenda for modernizing

government. On the face of it, modernization provides a new lease of life for the notion of effective government action informed by reason, a promise of a new 'post-ideological' approach to government, '. . . an approach where evidence would take the centre stage in the decision making process' (Davies *et al*. 1999, p. 3).

There are two main forms of evidence required in this approach to improving governmental effectiveness. The first is evidence to promote *accountability* in terms of results – evidence that government is working effectively. The second is evidence to promote *improvement* through more effective policies and programmes – evidence of how well such policies and programmes 'work' in different circumstances. These two forms of evidence are different in nature. The first is primarily in the form of *information* on attributes of performance and is reflected in the growth of performance management in government – the increasing use of performance indicators and targets, for example, in government departments' Public Service Agreements (PSAs) with the Treasury (HM Treasury 2000). The second form of evidence is qualitatively different from the first. Here we are talking about *knowledge* of how policy interventions achieve change in social systems. Conventionally, we assume that reliable knowledge provides a sound basis for effective action; it is explanatory and theoretical, providing an understanding of how policies work.

The focus of this paper is on this second form of evidence and the role of evaluation of public policies and programmes in helping to generate it and thereby improve policy making. Increasingly, we hear from government ministers that 'what matters is what works'; evidence of what works is to be provided through substantially increased research and evaluation programmes in government departments and greater use of pilot projects to test out new approaches (Martin and Sanderson 1999). Thus, according to Secretary of State Blunkett, again, research will '. . . help to determine what works and why, and what types of policy initiatives are likely to be most effective . . .' and '. . . must vastly improve the quality and sensitivity of the complex and often constrained decisions we, as politicians, have to make' (DfEE 2000a). The creation of the new Centre for Management and Policy Studies in the Cabinet Office was intended to provide a 'window at the heart of government' for research and evaluation evidence (Cabinet Office 2000).

The overall purpose of this paper is to explore the notion of evidence-based policy making in the context of policy intervention in increasingly complex social systems and the potential role of piloting and evaluation. The next section discusses the 'rationalist' basis of the notion of evidence-based policy making and critiques from constructivist and post-modernist perspectives. I then go on to discuss the implications of the growing appreciation of the complexity of social systems both for the quest to derive reliable social knowledge to guide policy action and for approaches to governance, arguing that a greater burden is placed upon policy experimentation,

evaluation and social learning. This is followed by a brief discussion of the UK Government's commitment to the piloting of new policies and programmes, in particular the limited scope provided for evaluation to derive sound evidence of whether policies work. The penultimate section argues that there is a need for greater clarity about the role of evaluation when such piloting essentially constitutes 'prototyping' and that more emphasis should be placed on developing a sound evidence base for policy through long-term impact evaluations of policies and programmes. I argue from a realist position that such evaluation should be 'theory-based' and focused on explaining and understanding how policies achieve their effects using multi-method approaches. The final section draws together conclusions and some issues requiring further attention.

RATIONALIST ASSUMPTIONS OF EVIDENCE-BASED POLICY MAKING AND MANAGEMENT

The notion of evidence-based policy making has gained renewed currency in the UK in the context of the current Labour Government's commitment to modernize government (Davies *et al*. 1999). In this vision of 'modern' government, policy making is more forward-looking, joined-up and strategic; public services are more responsive to the needs of users and are more efficient, effective and delivered to higher quality standards (Cabinet Office 1999a). A key driver of modernization is evidence-based policy making and service delivery – the new ministerial mantra is 'what matters is what works' (Cabinet Office 1999b). However, the concept of evidence-based policy making is rarely defined explicitly; one attempt to do so is provided by Plewis (2000, p. 96):

> New Labour proclaims the need for evidence-based policy, which we must take to mean that policy initiatives are to be supported by research evidence and that policies introduced on a trial basis are to be evaluated in as rigorous a way as possible.

Evidence can inform the development and implementation of policy in a number of ways. The emphasis is usually placed upon the two aspects included in Plewis' definition: first, evidence of the likely effectiveness of policy options to inform decisions on what policy action to take; and, second, evidence from evaluations of policies as implemented to inform decisions on whether to continue or how to adjust and improve policies and to contribute to the evidence base to inform future consideration of policy options. Rather less attention is given to two other important aspects of the policy process. The first of these is evidence of problems and needs requiring public policy intervention; a better understanding of the specific nature and incidence of social problems is fundamental to improving the effectiveness of policy responses. The second related aspect concerns the process of objective setting; thus, an improved understanding of the problem to be addressed and of the effectiveness of possible policy options will

help to inform the deliberations of key stakeholders in the process of setting objectives.

In the notion of modern government, evaluation in particular is required to play an enhanced role in providing information on performance to enhance accountability and 'control by results', and in providing evidence of what works to inform policy learning and improvement. However, this position is underpinned by certain assumptions about the nature of the policy-making process and the nature of valid social knowledge and these are contested, something which undermines the notion that there is one best way to improve evidence-based policy making. It is possible to see this renewed emphasis on evidence-based policy as a recourse to rationalism of a kind that has been seriously challenged, most vehemently in contemporary writing about post-modernity. On one level, the emphasis on evidence can be interpreted as simply contributing to the legitimation of policies and political commitments. According to Walker (2000, p. 62–3), research is but one influence on the policy process and '. . . is not always influential, supplanted by the powerful political forces of inertia, expediency, ideology and finance'. Kogan (1999) argues that governments will seek to legitimize their policies with reference to the notion of evidence-based decision making but use research evidence only when it supports politically-driven priorities. Cook (1997, p. 40) emphasizes that '. . . the politician's prime goal is to be re-elected rather than to respect technical evidence . . .' As J.M. Keynes once famously said: 'There is nothing a politician likes so little as to be well informed; it makes decision-making so complex and difficult.' (quoted in Davies *et al*. 1999, p. 3).

There is a long history of research on the relationship between social science and public policy making (Cook op. cit.; Weiss 1995a; Nutley and Webb 2000). The ideal model of evidence-based policy making is predicated upon certain assumptions relating to: the nature of knowledge and evidence; the way in which social systems and policies work; the ways in which evaluation can provide the evidence needed; the basis upon which we can identify successful or good practice; and the ways in which evaluation evidence is applied in improving policy and practice. Nutley and Webb (op. cit., p. 25) argue that the notion of evidence-based policy and practice '. . . fits well with a rational decision-making model of the policy process'. Such a conceptualization of the process of policy formulation and implementation has long dominated the field of policy studies (Colebatch 1998). Thus, it appears to be rational common sense to see policy as a purposive course of action in pursuit of objectives based upon careful assessment of alternative ways of achieving such objectives and effective implementation of the selected course of action. Moreover, rationality is enhanced by being clear about the objectives we wish to achieve and by evaluating the extent to which the policy as implemented actually achieves these objectives. If policy is goal-driven, evaluation should be goal-oriented.

Such evaluation completes the cycle and provides feedback to improve the policy (ibid.; Parsons 1995).

Within this rational model, then, the focus is on improving the 'instrumental' use of research and evaluation. Schwandt (1997, p. 74) argues that within this model of instrumental rationality, '. . . policymakers seek to manage economic and social affairs "rationally" in an apolitical, scientized manner such that social policy is more or less an exercise in social technology'. The ideal form of knowledge to provide a firm basis for this project is seen as derived through quantitative methodologies, empirically-tested and validated. Thus, Shaw (1999, p. 3) argues that the new enthusiasm for evidence-based policy gives a new lease of life to '. . . preoccupations with measurement, traditional worries regarding reliability and validity, and other concerns captured within quantitative methodology'. This perspective is illustrated by a recent study by the Performance and Innovation Unit of the UK Cabinet Office (2000) on improving the role of analysis in policy making which emphasizes the need for more and better data, better modelling, especially econometric, and more use of longitudinal and experimental research designs – an essentially quantitative agenda. Another Cabinet Office report on developing 'professional policy making' considers new skills needed by policy makers, amongst which are '. . . a grounding in economics, statistics and relevant scientific disciplines in order to act as "intelligent" customers for complex policy evidence' (Cabinet Office 1999b, para. 11.12).

This set of ideas about policy making maintains its influence in spite of sustained critique (Colebatch op. cit.). An important basis for criticism has been the constructivist or interpretivist position (Crotty 1998; Guba and Lincoln 1989), which argues that knowledge of the social world is socially constructed and culturally and historically contingent. From this perspective the notion of knowledge and its role in policy making becomes more complex; the quantitative agenda is seen as offering limited potential for improving the evidence base of policy making. It is argued that scientific knowledge can have no unique claim to objectivity, and that research does not simply inform policy development in an instrumental way but rather plays an important role in promoting broader 'enlightenment' of policy makers. From a constructivist perspective, policy development is seen as a '. . . process of deliberation which weighs beliefs, principles and actions under conditions of multiple frames for the interpretation and evaluation of the world' (Dryzek (1990) quoted in Van der Knaap 1995, p. 202). Policy learning involves a socially-conditioned discursive or argumentative process of development of cognitive schemes or frames which questions the goals and assumptions of policies. Evaluation research, and social science research generally, is seen as used more often in a conceptual rather than an instrumental way, reaching decision makers in unsystematic and diffuse forms, 'percolating' into the policy arena and influencing thinking about policy issues, providing policy makers with '. . . a background of infor-

mation and ideas that affected the ways that they viewed issues and the alternatives that they considered' (Weiss 1995a, p. 141).

The constructivist position in epistemology has generated a strong current of uncertainty, of contingency of thought and meaning, which is not sympathetic to traditional notions of evidence-based policy making. Indeed, this current has been strengthened by the work of postmodernist thinkers who, based upon a '. . . comprehensive dissatisfaction with the western humanist tradition. . .', have set '. . . a massive bonfire. . .' under modern rationalist assumptions (O'Sullivan 1993, p. 22). The postmodernist critique of reason rejects the notion of an objective, external world and the search for a 'metanarrative' to provide a secure foundation for knowledge. It is argued that there can be no certainties; everything is contingent upon a 'radicalised subjectivity'; our understanding of the world is a matter of interpretation through particular forms of thought and language in particular social and political contexts. According to Smart (1999, p. 63): 'We find ourselves abroad in a world in which social theory and analysis is no longer able, with any credibility, to provide a warrant for political practice and ethical decision-making.'

However, there is a real problem of reconciling these implications of the postmodernist position with the practical requirements of processes of collective decision making and action that rely on assumptions of 'grounded knowledge'. Oakley (2000, p. 298) argues that: 'The contention that there are no "objective realities" would obliterate the purpose of feminist (and other) emancipatory political projects'. Indeed, there is something of a paradox here. The conditions that have generated the postmodernist 'bonfire of the certainties' do not signal the end of the need for an analytical or cognitive basis for decision making and action. On the contrary, if anything they increase this need as the sense of the complexity of the social world is heightened. Thus, Smart (ibid.) argues that '. . . questions concerning political responsibility and ethical decision-making, the difficulties of adjudicating between the expression and pursuit of self-interest and the promotion and adequate provision of the public domain, as well as the problems encountered in everyday social life of making a choice or taking a stand, have if anything become analytically more significant. . .'.

COMPLEXITY AND SOCIAL LEARNING

The implications of the growing appreciation of the complexity of economic and social organization do indeed appear to be momentous. In Giddens' (1990) analysis of late modernity, the conditions of 'wholesale reflexivity' undermine the notion of certainty in relation to social knowledge and the notion that '. . . more knowledge about social life . . . equals greater control over our fate. . .' (ibid., p. 43). On the contrary, knowledge of the social world actually contributes to its instability, to conditions of uncertainty and ambivalence. There will always be unintended and unanticipated consequences of action which undermine our capacity to predict and control on

the basis of knowledge. But this does not imply that systematic knowledge of human action is not possible; it implies that there are no 'foundationalist' guarantees but nevertheless that our knowledge of the social world can help us '. . . in our attempts to steer the juggernaut' (ibid., p. 154).

Of course, epistemological dispute continues about the basis upon which 'reliable' knowledge of the social world can be derived, but the realist tradition in the philosophy of social science is enjoying something of a revival in the context of the concern to make policy making more evidence based (Archer 1995; Archer *et al*. 1998; Kirk 1999; Pawson and Tilley 1997; Searle 1995; Trigg 2001). Realists argue that there are social phenomena independent of cognition to be explained in terms of underlying mechanisms (which may not be directly observable) and that the task of social science is to understand the way in which mechanisms work in conjunction with contextual factors to generate social outcomes. As Trigg (2001, p. 237) argues from this standpoint, '. . . there is little point in furthering social science if it is useless in helping to deal with the human world in which we are situated'. According to Pawson and Tilley (op. cit.), the task is to understand what works, for whom, in what circumstances, and why as a basis for piecemeal social reform; indeed, the phrase 'what matters is what works' has become something of a mantra in evidence-based policy circles. Realists argue that they provide the basis for a 'middle ground' between the over-optimistic claims of objectivists on the one hand and over-pessimistic nihilism of relativists on the other (Trigg 2001).

Realism therefore offers the prospect of 'steering the juggernaut' on the basis of a better understanding of what is likely to work in terms of public policies and programmes. This provides a potentially important basis for effective governance but a broader institutional framework is required to deal with social complexity that goes beyond traditional command and control models (Mulgan 1998). Amin and Hausner (1997) have developed the notion of 'interactive governance' as the basis for strategic guidance of increasingly complex societies. They argue that the idea of society as a web of interlocking networks of affiliation and interaction, structured around a multiplicity of formal and informal institutions, constitutes '. . . a powerful metaphor for grasping the problems of social complexity' (ibid., p. 10). Networking (or 'relational interaction'), involving both state and non-state governance structures, provides a basis for overcoming the rigidities associated with hierarchy – interactive, deliberative networks with a multiplicity of shared values and responsibilities being more discursive and democratic (ibid., pp. 14–19). Strategic guidance – the ability to co-ordinate, arbitrate and facilitate multiple governance networks – is seen as 'the quintessence' of governing social complexity (ibid., p. 18).

Jessop (1997, p. 111) argues that a key element of such strategic guidance is '. . . reflexive monitoring and dynamic social learning. . .'. Other commentators have recognized the need for an enhanced capacity for learning as a means of reconciling the implications of increasing social com-

plexity with the requirements of effective public policy intervention. Donald Campbell's notion of the 'experimenting society' was founded upon a commitment to innovation, social reality-testing and learning (Campbell and Russo 1999). As Oakley (2000, p. 320) states:

> The experimenting society would be active, preferring innovation to inaction; it would be an evolutionary, learning society and an honest society, committed to reality testing, to self-criticism and to avoiding self-deception; it would be non-dogmatic, accountable, decentralized and scientific. . ..

Some years ago Dunsire (1986) argued that under conditions of uncertainty about the ex ante 'correctness' of policy decisions, and about capacities to implement policies as intended, there is a need to strengthen the role of evaluation in providing up-to-date, relevant information on actual performance, and to build the capacity to take action to modify policy design and implementation in the light of such information. The implications of Rescher's (1998) analysis of complexity point towards increasing need for monitoring and evaluation of 'how matters work themselves out':

> The fact is that in situations of unmanageable complexity, practice in matters of public policy is often guided more effectively by localized experimental trial-and-error than by the theorizing resources of an intellectual technology unable to cope with the intricacy of interaction feedbacks and impredictable (sic) effects. (ibid., p. 189)

In this situation, a major burden is placed upon policy experimentation and evaluation as key institutional practices in interactive governance to provide the basis for reflexive social learning. With a realist commitment in policy and programme evaluation to expanding the evidence base on 'what works', coupled with governance processes that embody a serious commitment to learning from such evaluation, we can potentially achieve a rational basis for the guidance of social change towards collectively desired ends. On the face of it, at least, the UK government's commitment to the piloting of new policy developments, accompanied by increasing emphasis on evaluation as a key basis for evidence-based policy making, appears to be consistent with this model of interactive governance.

THE CONTRIBUTION OF PILOTING TO EVIDENCE-BASED POLICY MAKING

The scale of piloting and testing of new policy developments by the New Labour Government in the UK has indeed been significant in areas such as crime prevention, employment and welfare policy, health, education and local government. Employment and welfare policy has been a significant area for piloting given its importance to the government's Welfare-to-Work agenda. Thus, the New Deals which have been developed for various groups experiencing particular disadvantage in the labour market (e.g.

young people, lone parents, disabled, long-term unemployed) were all subject to initial piloting or 'prototyping' before being extended on a national basis (Hasluck 2000). The Single Work-Focused Gateway for welfare benefits (later branded ONE), designed to provide a single point of entry to the benefits system for those of working age but not in full-time work, was piloted in 12 areas across Great Britain starting in June or November 1999 (Chitty 2000). Health Action Zones (HAZs) were developed as pilot projects designed to explore mechanisms for breaking through current organizational boundaries to tackle inequalities and deliver better health service and health care, encouraging co-operation across the National Health Service. Eleven HAZs were established in April 1998 as 'trailblazers' and 15 more were set up in April 1999 (Judge, *et al*. 1999).

In relation to education policy, the 'Sure Start' initiative was developed as a key element of the government's drive to tackle child poverty and long-term social exclusion. Its aim is to develop improved and co-ordinated local services, owned by the local community, for families, which will enable children to thrive when they are ready to enter primary school. During 1999–2000, 126 'trailblazer' programmes were established (Sure Start Unit 2000). A final example of piloting comes from the UK government's agenda for the modernization of local government, a key element of which has involved the introduction of a new statutory duty of Best Value to replace the previous regime of Compulsory Competitive Tendering (CCT). This duty requires local authorities to secure continuous improvement in their performance defined in terms of an extensive set of performance indicators and designated targets and standards. Authorities are required to undertake performance reviews of all services over a five-year cycle and produce an annual Local Performance Plan for local people. There is also a rigorous regime of audit and inspection, backed by new powers, allowing government intervention in the event of 'clear performance failure'. Pilot projects were established in 40 local authorities and two police forces in England (with a separate piloting exercise in Wales) commencing in April 1998, which ran for over two years (Martin and Sanderson 1999).

All of these pilot programmes have been (or are being) subject to comprehensive evaluation. Although they have obviously differed in specific terms, nevertheless they have all involved a combination of an assessment of impact with the analysis of implementation processes. In conventional terms, they have sought to combine summative and formative evaluation. At least at the stage of developing the evaluation design, the ostensible intention has been to provide policy makers with feedback both on outcomes, effects and impacts achieved (and, in most cases, value for money) and on the effectiveness of approaches to delivery of the programmes and lessons in terms of 'good practice'. On this basis, therefore, evaluations have been set up to answer two key questions: first, 'does it work?'; and, second, 'how can we best make it work?'

The relative balance of emphasis that can be placed upon these two ques-

tions depends, in turn, on two key factors: first, the degree of weight or importance attached to them by policy makers; and, second, the extent to which sound evidence can be obtained as a basis for providing a convincing answer. Taking the latter issue first, it is clear that there have been significant limitations on the ability of pilot evaluations to provide convincing evidence to answer the question 'does it work?' Thus, there are a number of problems that limit the scope for impact evaluation in the context of pilot programmes. The first problem concerns the time needed for the effects of new policies to be manifested and to become capable of measuring and isolating from other factors, particularly where policies are seeking to tackle deep-seated economic and social problems. It may take some considerable time for pilot projects to become fully established so as to represent the conditions under which a policy would work when fully implemented. If the policy aims to change attitudes and behaviour or achieve institutional reform, effects may be difficult to achieve during the course of a pilot project.

This problem is exacerbated by political considerations that constrain the length of time for pilot programmes to operate. When policy initiatives arise from political manifesto commitments, policy makers are understandably impatient to receive results that will provide evidential support for decisions to proceed with full implementation. Such a political interest potentially conflicts with the interests of evaluation research, the interests of which are served by long-term, in-depth analysis of the effects of pilots. As Walker (op cit., p. 162) argues, '. . . the short-term horizons of policy making will outpace those of research. Political considerations, with an upper and lower case "p", will continue to have precedence over those to do with research design.' Therefore, on the one hand, the time made available for piloting may be insufficient for the analysis of impact. On the other hand, even within the timescale of the pilot, as Chitty (op cit., pp. 9–10) emphasizes, policy makers and politicians will be looking for results as early as possible to enable them to 'plan the way ahead'.

Related to the problem of timescale, a second set of difficulties arises in seeking to isolate effects of pilot programmes from exogenous change and from the effects of other initiatives that may also be having an impact on the same problems as those addressed by the pilot. There are two key problems here. First, the size of the impact may not be substantial and therefore may be difficult to measure at an acceptable confidence level. For example, from a cross-national comparison of the effectiveness of welfare-to-work polices, White (1999, p. 69) concludes that the effects, though positive, '. . . are generally moderate rather than large'. The second problem is that of 'attribution' – given observed and measured outcomes, how can the effect of the piloted measures be isolated from other influences. This is especially difficult in the context of area-based initiatives (ABIs), which are targeted on particular areas of greatest need, and which are playing a particularly important role in the UK government's polices to address social

exclusion (DETR 2000a). ABIs that are being (or have been) piloted include Health Action Zones, Education Action Zones, Sure Start and Employment Action Zones. A key feature of these policies is the aim of joining up programmes and mainstream services through holistic approaches on the assumption that the whole will be more than the sum of the parts. In such a situation it is extremely difficult to isolate the effects of any one initiative and the effect due to interaction with other initiatives and with mainstream services.

In such a context, evaluators face severe difficulties in seeking to isolate policy and programme impacts. In particular, the feasibility and validity of the experimental design, involving randomized controls to isolate net additional impact (i.e. net of deadweight effects), is undermined because the *ceteris paribus* requirements cannot be fully satisfied. In relation to ABIs, the range and variety of policy initiatives and the scale of change in local government renders strict control impossible (Granger 1998). In national initiatives, such as the New Deal programmes, the use of control groups is, as Hasluck (op cit., p. 63) argues, not practical primarily due to ethical objections to denying some eligible people the benefits of the initiative. In such situations, a range of quasi-experimental approaches can be pursued (as in the New Deal evaluations) but these will not provide the kind of unequivocal results that policy makers seek.

There are a number of other problems that add to the complexity of evaluation and create difficulties for impact evaluation. In the situation of piloting there is considerable scope for variation in the implementation of a policy; such variation may indeed be encouraged to generate different approaches for the purposes of evaluation. Thus, there is also likely to be a lack of stability over time in the form of intervention. Indeed, these problems are to some degree built in to pilots when results of evaluations of implementation process provide the basis for ongoing improvement and sharing of good practice. Moreover, where the aim is to tailor help to the specific needs of individuals or groups (as is increasingly the case in, for example, welfare-to-work initiatives), it becomes difficult to define a discrete, standard intervention of the kind required by experimental designs. Finally, there are a number of other reasons why a pilot may not be typical of the policy as it would ultimately be implemented. For example, as Hasluck (op cit., p. 63) points out, '. . . the resources devoted to the pilot may exceed that (sic) available at national roll out. There may also be a greater level of commitment and a "pioneering spirit" amongst staff involved in delivery'.

Indeed, it could be argued that there is a 'structural' danger of unrepresentativeness of pilots in a context where there is a strong political commitment to a policy, and the pilot receives generous resourcing in order to make it work. This brings us back to the question posed earlier regarding the relative weight given by policy makers to knowing whether a policy works as opposed to how it can be made to work. We have seen that there

are problems with the evidential basis for answering the first of these questions. There are also grounds for arguing that policy makers and politicians have also placed more weight on the second question. I have already referred to the dominance of short-term political considerations and Walker (op cit., p. 161) suggests that '. . . the use of pilots has been more akin to prototyping than to experimentation. . .'. An example is provided by the Best Value pilot programme in Great Britain. As Martin and Sanderson (op cit.) argue, the relative emphasis on summative and formative evaluation shifted significantly towards the latter during the course of the pilot programme. Early in the programme, in advance of any substantial evaluation results, the government published a series of key policy documents outlining the way in which it expected the Best Value framework to operate. Martin and Sanderson conclude (ibid., p. 254) that:

> it is . . . clear that, at this stage in the policy and electoral cycle at least, central government is not primarily interested in or particularly receptive to alternatives to Best Value. The Best Value principles are not up for debate – ministerial advisers claim to have known what was required before Labour came to power. Piloting is not therefore designed to test whether the Best Value framework should replace CCT – ministers are already committed to this course of action. There is some scope for 'single loop learning' focused on ways of making the framework operate more effectively at local level. However, the piloting process is not so much about experimenting as about exemplifying. The main aim seems to be to create a cadre of 'trail blazing' authorities that will provide instances of 'good practice' that can be adopted by other councils.

SOME IMPLICATIONS FOR EVALUATION

There are grounds, therefore, for questioning whether the government's commitment to evidence-based policy making extends to using piloting as genuine experimentation to obtain evidence about whether policies work. The endemic problems that exist in seeking to obtain sound and convincing evidence of the impact of piloted policy initiatives, together with the uncertain role of research and evaluation in informing key policy commitments, raise doubts about the role of pilot programme evaluation. If policy makers are not able to create the conditions for piloting that would enable robust impact evaluation to be undertaken (as occurs much more in the USA – see Riccio 2000), then we might legitimately question whether it is worth persisting with the notion that pilots can genuinely inform decisions on *whether* to proceed with a policy.

Consequently, there is a need for greater clarity about the role of piloting and, therefore, the purpose of evaluation. In particular, there is a need to be clear when piloting is primarily about 'prototyping' in which '. . . greater emphasis is placed on *how* it works than *whether* it works . . .' and '. . . evaluation is as much about identifying good practice as about identifying the "counterfactual". . .' (Chitty 2000, op cit., p. 13). In this context,

the form of evaluation required is close to action research, with a focus on identifying how implementation can be improved, working closely with practitioners. However, we are then left with the problem of how to obtain better evidence about whether policies work and about what works and why in different circumstances. Can piloting make a significant contribution in this respect?

This would require genuine policy and programme pilot experiments in which considerations of evaluation design played a major role, so as to create the conditions under which robust evidence could be obtained. Such evidence is required both about the nature of effects and impacts of policies and about *how* these effects are produced in the circumstances in which they are implemented (where relevant, in combination with other policy measures). However, the prospects of developing such an approach to piloting in the UK policy context is uncertain given the rather short-term horizons of policy making and a political culture that has long been unsympathetic to 'rational knowledge'. In this respect, there is a stronger tradition of experimentation in the USA, which has permitted evaluators to properly assess the effectiveness of social programmes in pilot situations (Riccio 2000, p. 1).

In the UK context, therefore, it may be that the most feasible strategy for strengthening evidence-based policy making comprises two related elements. First, as Chitty (op cit., p. 12) argues, there is a need to ensure that all the currently available relevant research and evaluation evidence is thoroughly reviewed and synthesized and used to inform policy thinking and appraisal. Indeed, there have been significant developments in this direction, and UK government departments and agencies are placing more emphasis on such review and synthesis work. A recent example is a review commissioned by the DETR of the evidence base for regeneration policy and practice with a view to '. . . ensuring that policy and practice are informed through a more rigorous and challenging evidence base' (DETR 2000c). A key development was the establishment in 1992 of the Cochrane Foundation in the health field, the aim of which is '. . . to help people make well-informed decisions about healthcare by preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare interventions' (Cochrane Collaboration 1999, p. 1). The UK Centre was established in 1992 with support from the National Health Service (NHS). In 1999, the Campbell Collaboration was established to emulate the Cochrane Collaboration in the fields of social and educational policy (Boruch, *et al*. 1999). The Department for Education and Employment (DfEE) has recognized the importance of review and synthesis of research evidence and has expanded its research budget to fund dedicated research centres as '. . . foci for the synthesis of knowledge as well as adding to it. . .' in relation to key topics in education and employment (DfEE 2000b).

The Cochrane Collaboration is regarded as a success but a number of important issues and questions have been raised about the potential for

successful research synthesis and review in areas of social policy (Oakley 1999). The key difficulties are those that bedevil impact evaluation in the context of complex social initiatives; some of these were discussed in the previous section. Thus, many social interventions comprise packages of approaches, with component elements 'treating' different groups, and interventions classified together may comprise different specific measures. A focus on outcomes '. . . might lose much of the variety and detail about interventions in which practitioners and policy makers are most interested, and which they view as critical in distinguishing between different approaches' (ibid., p. 4). Moreover, many social interventions are explicitly targeted at multiple outcomes. Finally, reviews may have difficulty addressing underpinning theory '. . . unless review activity is structured to cross problem/outcome areas, and allow for the classification of interventions according to their theoretical base' (ibid., p. 5).

These difficulties undermine the potential to derive clear messages about what works from reviews and syntheses of research and evaluation studies of social policy interventions. The last issue referred to above, relating to the theoretical basis of policy intervention, is of particular concern from a realist perspective, where the key focus is on identifying and explaining what works in terms of underlying policy 'mechanisms' operating in particular contextual circumstances. Thus, Pawson (2000) criticizes dominant approaches to research synthesis, which he labels 'numerical meta-analysis' and 'narrative review', with their concern to identify 'best buys' for policy makers. He argues that they do not achieve decisive results, the former making no effort to understand how programmes work in different contexts, the latter unable to provide transferable lessons. Pawson argues for a realist approach to synthesis that focuses on families of mechanisms and produces tailored, transferable theory.

The next few years should provide some important lessons on the value of research synthesis and review in terms of providing policy makers with a better understanding of what works, for whom, in what circumstances in areas of social policy. At the present time, there are grounds for concern about the nature of research and evaluation evidence available in many areas. From a realist perspective, much of the available evidence fails to provide the basis for a theoretically-grounded understanding of transferable lessons about what works and why (Pawson, op cit.).

This brings us to the second element of the strategy for improving the evidential basis of policy making: better generation of new evidence through long-term evaluation of new policy initiatives once they are implemented (after any prototyping or piloting). Such evaluation needs to be planned jointly with implementation in order to provide the best possible conditions for detailed longitudinal research from a well-defined baseline position.

Some recent initiatives by the UK government to put in place long-term evaluations of key policy initiatives offer some hope for this strategy, with

a strong emphasis on building the evidence base to improve the effectiveness of future policy intervention. Two such evaluations will be very briefly considered, the first relating to the New Deal for Communities programme and the second to the Best Value regime for local government, the pilot phase of which was discussed above.

The New Deal for Communities (NDC) is a key programme in the government's National Strategy for Neighbourhood Renewal, the development of which was led by the Social Exclusion Unit (Cabinet Office 1998). Launched in September 1998, NDC is focused on small areas (1000–4000 households) suffering from the worst problems of deprivation and social exclusion. It aims to close the gap between them and the rest of Britain in terms of worklessness, health, crime, educational achievement and broader economic prosperity and quality of life. Bids were developed by inclusive local partnerships, which were required to '. . . harness the active involvement of local communities' (DETR 1999, p. 1). Successful partnerships developed long-term (10-year) strategies for their neighbourhoods that built on existing services and programmes to produce '. . . joined up solutions to joined up problems . . .' (ibid.). By May 2001, 39 partnerships had been approved, comprising 17 'pathfinders' and 22 'second round' partnerships. Of the total of 39 partnerships, nine are in London boroughs and 17 are in metropolitan local authorities.

Considerable effort and resources are being put into evaluation of NDC by the government. Local partnerships are required to undertake their own evaluation work and this is complemented by a national evaluation. The Department for the Environment, Transport and the Regions (DETR) commissioned a 'development project' to work up proposals and recommendations for the national evaluation in view of its scale and potential complexity. A key feature is its long-term nature, potentially up to the whole 10 years of the programme, thus providing scope for robust analysis of impact. The DETR (re-designated the Department of Transport, Local Government and the Regions – DTLR – following the 2001 General Election) saw the evaluation as comprising much that is 'conventional' in the regeneration evaluation world: the analysis of outcomes and impacts, addressing sustainability and additionality; addressing the problem of attribution, potentially through comparison areas; the analysis of implementation processes; the assessment of value for money; and the provision of feedback on 'quick wins' and early lessons (ibid.).

However, there are some aspects of the evaluation, receiving particularly strong emphasis, that require more innovative development of evaluation approaches and methodologies. First, the national evaluation was required to 'work alongside and support' local partnerships and avoid a heavy, top-down approach; indeed, the DETR considered 'action research' to be a key element in the approach to help partnerships to improve their effectiveness through useful and timely feedback. Second, given the emphasis in NDC on 'joining up', there was a need for robust approaches to analysing 'syn-

ergy' effects: how NDC links with and adds value to other area-based initiatives and mainstream programmes and services; and how the latter are re-oriented over time in a sustainable way better to meet the needs of the locality. Third, the evaluation was required to identify 'what works and how it works' and to strengthen the evidence base about effective practice in tackling a wide range of problems at the neighbourhood level. Given the complexities of NDC strategies, there is considerable potential in this evaluation to make a significant contribution to the development of evaluation approaches and methods, especially in the realm of theory-based evaluation employing multi-method designs.

Turning to the second example, I argued above that the pilot phase of the introduction of the Best Value regime in local government in England and Wales provided a good case of 'prototyping', with the emphasis more on learning how to implement the regime effectively than on understanding whether and how it worked. However, with the introduction of the new regime throughout local government from April 2000, the opportunity arose for a long-term evaluation of its impact in the context of the broader 'modernization agenda' for local government. The DETR indeed commissioned such an evaluation, which is analysing over a five-year period the way in which authorities have implemented Best Value processes, the effects of such processes through organizational and cultural change, and the impact of the changes on the performance of authorities (DETR 2000b).

Again, this is a complex evaluation, presenting some significant design and methodological challenges in terms of capturing and measuring key changes; attributing such changes to different aspects of Best Value as implemented in different socio-economic, policy and organizational contexts (together with other aspects of the local government modernization agenda); and understanding how the changes are brought about, feeding back lessons on '. . . which mechanisms . . . appear to be most suitable and effective in particular circumstances and why' (ibid., p. 13). Therefore, as with the NDC evaluation, there is a strong emphasis on developing the evidence base about effective practice. Within a longitudinal design, the need for a mix of quantitative and qualitative methods was emphasized. These include multi-variate statistical analysis, using local authority service performance data and survey data to identify relationships between processes, contextual factors, organizational and cultural changes and performance outcomes, and in-depth longitudinal case study research to understand how and why (or why not) process and organizational changes have worked to produce improved performance.

The need to develop such an understanding and make a real contribution to the evidence base for policy development presents some major challenges for evaluation. It is beyond the scope of this paper to discuss these in detail (see Sanderson 2000) but some key points can be addressed briefly. First, I would argue that evaluation must move beyond its traditional concern with measuring effect sizes and degrees of goal achievement to

embrace a theory-based approach to explanation. As Chen (1990) and Weiss (1995) have argued, social programmes are based upon explicit or implicit theories about how and why the programme will work and the task of evaluation is to surface those theories, identify the key assumptions and test their validity.

However, theory-based evaluation, while holding out considerable promise for the development of knowledge to underpin more effective policy making, nevertheless presents significant challenges in terms of articulating theories, measuring changes and effects, developing appropriate tests of theoretical assumptions, and in terms of the generalizability of results obtained in particular contexts. A key challenge is provided by the problem of causal attribution, which is central to theory testing (Chen 1990). Traditionally, evaluations have sought to isolate effects by establishing the counterfactual and measuring deadweight (what would have happened in the absence of the intervention); these have tended to be seen essentially as problems of measurement and estimation. However, theory-based evaluation requires an understanding of how effects have been produced and, particularly from a realist perspective, this requires an analysis of the mechanisms at work (Pawson and Tilley 1997). In the context of complex, multifaceted social interventions, Granger (1998) argues that it is 'virtually impossible' to achieve precise and clear-cut causal attribution and that we need clear, strong theories as a basis for counterfactual reasoning and causal inference, which at best may achieve a modest standard.

In the context of policy responses to complex economic and social problems, it is now widely argued that the best hope for 'generating trustworthy causal inferences', as Granger (ibid.) puts it, is through mixed methodology or multi-method evaluation designs. Such an argument, in fact, has a sound pedigree in the work over the years of Donald Campbell who, although associated with experimental evaluation, nevertheless maintained a critical, post-positivist position from which he argued, for example, that:

> experimental research is equivocal and ambiguous in its relation both to the real physical process involved and to scientific theory. This equivocality calls for use of multiple methods, none of them definitional, triangulating on causal processes that are imperfectly exemplified in our experimental treatments and measurement processes. (Campbell and Russo 1999, pp. 132–3)

As Riccio (op cit.) argues in the context of welfare-to-work and employment evaluations, since randomized social experiments are '. . . often too blunt a method for testing important theories or hypotheses underlying a particular social program. . .' (ibid., p. 1), '. . . new efforts to augment social experiments with nonexperimental and quasi-experimental strategies . . . hold promise for improving our capacity to understand the efficacy of social initiatives, and for assessing, with greater scientific rigor, the validity of

theoretical assumptions upon which they rest' (ibid., p. 16). This trend is also highlighted by White (1999, p. 69) in the UK context:

> (T)o learn more about welfare-to-work programmes will require greater co-operation between what have in the past been distinct approaches to evaluation. Rigorous quantitative evaluation will continue to be important, but so will the use of qualitative, case-study and organisational research. Only if this takes place will we be able to learn not only what works better, but how it works better.

CONCLUSION

The resurgence of interest in evidence-based policy making indicates the continuing force of optimism about the potential to achieve social progress through the application of reason. This paper has sought to present a case for evaluation as playing an increasingly important role in this project, but evaluation conceived somewhat differently than in traditional accounts. A focus on the role of evaluation in reflexive policy learning is required to resolve a paradox in late modern society: that while increasing complexity of social systems progressively undermines notions of certainty in social knowledge it simultaneously raises the stakes in relation to rational guidance of those systems.

Consideration of the implications of complexity raises our awareness of the limits to prediction and control of non-equilibrium social systems and the increasing significance of unintended consequences of our actions. While we can retain some confidence in our ability to understand and explain the behaviour of such systems, this needs to be tempered with a degree of modesty about what we can achieve. Thus, we need to recognize that policies are essentially 'conjectures' based upon the best available evidence. In most areas of economic and social policy this evidence will provide only partial confidence that policy interventions will work as intended. Consequently, such policy conjectures must be subject to rigorous testing. Evaluation is required to assess and understand how policies have worked (or not) and why, so that lessons can be learned to inform improvements. I have argued that policy making and evaluation need to be conceived as instances of 'practical reason' rather than as solely technical exercises. If evaluation is to fulfil its potential for driving policy learning, it must be fully integrated into the ongoing discourse, able to sustain advocacy of the 'evidential voice' and help policy makers to '. . . think more intelligently about the domain in which they worked' (Weiss, op cit., p. 141).

A key requirement of this model of policy learning, then, is the strengthening of evaluation as an explanatory enterprise. I have referred to the resurgence of interest in realist philosophical stances, which pose the key evaluation question in explanatory terms: 'what works for whom under what circumstances, and why?' The key challenge lies in allying this to complexity in the context of non-equilibrium social systems where the isolation of simple context-mechanism-outcome configurations is limited in

terms of both feasibility and usefulness. Again, as Schmid (1996, p. 205) has autioned, '. . . some deliberate eclecticism and modesty. . .' may be required with greater emphasis on developing a sounder descriptive base to explanatory efforts through, for example, policy mapping and pattern recognition approaches to identifying emergent properties of systems.

Such a note of caution should be taken seriously. To realize the promise of theory-based evaluation in the context of policy initiatives to address complex economic and social problems represents a major challenge, especially where integrated policy solutions are sought to address problems in joined up ways. Some implications for the practice of evaluation have been discussed but in conclusion it is possible to point up two issues requiring further attention. The first concerns notions of 'theory' that are relevant in the pursuit of theory-based evaluation in this context. I have referred to the argument that we may have to accept modest standards of causal inference and this raises questions about the forms of theoretical knowledge that can be derived and about how 'robust' such knowledge can be.

The second related issue concerns the role of better analysis and understanding of social problems and needs in improving the effectiveness of policy responses. As argued earlier, this receives somewhat less attention in the evidence-based policy debate than does the role of evaluation. However, such 'fundamental' research can be seen as complementary to better theory-based evaluation, providing the basis for clearer and more specific elaboration of hypotheses for testing in evaluation research and thus helping to offset the limitations discussed above relating to standards of causal inference. Of course, such fundamental research has been the focus of substantial support over the years, for example from the ESRC and Joseph Rowntree Foundation, but the argument here implies the need for closer links between such research and government-sponsored evaluations that are long term, theory-based and designed to build the evidence base of what works in policy terms.

REFERENCES

Amin, A. and J. Hausner. 1997. *Beyond market and hierarchy: interactive governance and social complexity*. Cheltenham: Edward Elgar.

Archer, M.S. 1995. *Realist social theory: the morphogenetic approach*. Cambridge: Cambridge University Press.

Archer, M.S., R. Bhaskar, A. Collier *et al*. 1998. *Critical realism: essential readings*. London: Routledge.

Bernstein, R. 1991. *The new constellation: the ethical-political horizons of modernity/postmodernity*. Cambridge: Polity Press.

Boruch, R., A. Petrosino and I. Chalmers. 1999. *The Campbell collaboration: a proposal for systematic, multinational and continuous reviews of evidence*, Campbell Collaboration Meeting, University College, London, 15/16 July.

Bronk, R. 1998. *Progress and the invisible hand: the philosophy and economics of human advance.* London: Warner Books.

Bulmer, M. (ed.) 1987. *Social science research and government: comparative essays on Britain and the United States*. Cambridge: Cambridge University Press.

Cabinet Office. 1998. *Bringing Britain together: a national strategy for neighbourhood renewal*. Cm 4045. London: The Stationery Office.

——. 1999a *Modernising government*. Cmnd. 4320. London: The Stationery Office.
——. 1999b *Professional policy making for the twenty first century*. Strategic Policy Team. London: Cabinet Office.
——. 2000. *Adding it up: improving analysis and modelling in central government*. London: Central Office of Information.
Campbell, D.T. and M.J. Russo. 1999. *Social experimentation*. Sage Classics 1. Thousand Oaks: Sage Publications.
Chen, H.-T. 1990. *Theory-driven evaluations*. Newbury Park: Sage Publications.
Chitty, C. 2000. 'Why Pilot?' Paper presented to Royal Statistical Society Conference: *The evaluation of economic and social policies*, RSS, London, 4th July.
Cochrane Collaboration. 1999. *The Cochrane collaboration: a brief introduction*, Website: http://cochrane.org.
Colebatch, H.K. 1998. *Policy*. Buckingham: Open University Press.
Cook, T.D. 1997. 'Lessons learned in evaluation over the past 25 years', in E. Chelimsky and W.R. Shadish (eds), *Evaluation for the 21st century: a handbook*. Thousand Oaks: Sage.
Crotty, M. 1998. *The foundations of social research: meaning and perspective in the research process*. London: Sage Publications.
Davies, H.T.O., S.M. Nutley and P.C. Smith. 1999. 'Editorial: What works? The role of evidence in public sector policy and practice', *Public Money and Management*, Vol. 19, No. 1, 3–5.
DETR. 1999. *New deal for communities: an overview*. London: Department for the Environment, Transport and Regions.
——. 2000a. *Joining it up locally*. Report of Policy Action Team 17, National Strategy for Neighbourhood Renewal, London: Department for the Environment, Transport and Regions.
——. 2000b. *The long-term evaluation of best value and its impact*. Research Project Specification. London: Department for the Environment, Transport and Regions.
——. 2000c. *A review of the evidence base for regeneration policy and practice*. London: Department for the Environment, Transport and Regions.
DfEE. 2000a. *Influence or irrelevance: can social science improve government?* London: Department for Education and Employment.
——. 2000b. *Research programme 2000–01*. London: Department for Education and Employment.
Dryzek, J.S. 1990. *Discursive democracy: politics, policy and political science*. Cambridge: Cambridge University Press.
Dunsire, A. 1986. 'A cybernetic view of guidance, control and evaluation in the public sector', in F.-X. Kaufmann, G. Majone, V. Ostrom and W. Wirth (eds), *Guidance, control and evaluation in the public sector*. Berlin: Walter de Gruyter.
Giddens, A. 1990. *The consequences of modernity*. Cambridge: Polity Press.
Granger, R.C. 1998. 'Establishing causality in evaluations of comprehensive community initiatives', in K. Fulbright-Anderson, A.C. Kubisch and J.P. Connell (eds), *New approaches to evaluating community initiatives, volume 2: Theory, measurement and analysis*. Washington DC.: Aspen Institute.
Guba, E.G. and Y.S. Lincoln. 1989. *Fourth generation evaluation*. Newbury Park: Sage.
Hasluck, C. 2000. *Early lessons from the evaluation of the new deal programmes*, Research Report No. 49, Sheffield: Employment Service.
HM Treasury. 2000. *2000 spending review: public service agreements*. Cm. 4808. London: HM Treasury.
Jessop, B. 1997. 'The governance of complexity and the complexity of governance: preliminary remarks on some problems and limits of economic guidance', in A. Amin and J. Hausner (eds), *Beyond Market and Hierarchy: Interactive Governance and Social Complexity*. Cheltenham: Edward Elgar.
Judge, K. *et al*. 1999. *Health action zones: learning to make a difference*. Canterbury: Public and Social Services Research Unit, University of Kent.
Kirk, R. 1999. *Relativism and reality: a contemporary introduction*. London: Routledge.
Kogan, M. 1999. 'The impact of research on policy', in F. Coffield (ed.), *Research and policy in lifelong learning*. Bristol: Policy Press.
Martin, S. and I. Sanderson. 1999. 'Evaluating public policy experiments: measuring outcomes, monitoring progress or managing pilots?' *Evaluation*, Vol. 5, No. 3, 245–58.
Mulgan, G. 1998. *Connexity: responsibility, freedom, business and power in the new century*. London: Vintage.
Newby, H. 1993. *Social science and public policy*. Swindon: Economic and Social Research Council.

Nutley, S. and J. Webb. 2000. 'Evidence and the policy process', in H.T.O. Davies, S.M. Nutley and P. Smith (eds), *What works? Evidence-based policy and practice in public services*. Bristol: Policy Press.

Oakley, A. 1999. *An infrastructure for assessing social and educational interventions: the same or different?* Campbell Collaboration Meeting, University College, London, 15/16 July.

——. 2000. *Experiments in knowing: gender and method in the social sciences*. Cambridge: Polity Press.

OECD. 1994. *Performance management in government: performance measurement and results-oriented management*, Public Management Occasional Papers, No. 3, Paris: OECD.

O'Sullivan, N. 1993. 'Political integration, the limited state, and the philosophy of postmodernism', *Political Studies*, Vol. XLI, 21–42.

Parsons, W. 1995. *Public policy: an introduction to the theory and practice of policy analysis*. Cheltenham: Edward Elgar.

Pawson, R. 2000. *A realist approach to meta-analysis and research review*, Paper presented to UK Evaluation Society Annual Conference, Church House, London, 7/8 December.

Pawson, R. and Tilley, N. 1997. *Realistic evaluation*, London: Sage Publications.

Plewis, I. 2000. 'Educational inequalities and education action zones', in C. Pantazis and D. Gordon (eds), *Tackling inequalities: where are we now and what can be done*. Bristol: Policy Press.

Rescher, N. 1998. *Complexity: a philosophical overview*. New Brunswick: Transaction Publishers.

Riccio, J.A. 2000. 'Extending the reach of randomized social experiments: new directions in evaluations of American welfare-to-work and employment initiatives', Paper presented to Royal Statistical Society Conference: *The evaluation of economic and social policies*, RSS, London, 4th July.

Sanderson, I. 2000. 'Evaluation in complex policy systems'. *Evaluation*, Vol. 6, No. 4, 433–54.

Schmid, G. 1996. 'Process evaluation: policy formation and implementation', in G. Schmid, J. O'Reilly and K. Schomann (eds), *International Handbook of Labour Market Policy and Evaluation*. Cheltenham: Edward Elgar.

Schwandt, T.A. 1997. 'Evaluation as practical hermeneutics'. *Evaluation*, Vol. 3, No. 1, 69–83.

Searle, J.R. 1995. *The construction of social reality*. London: Penguin Books.

Shaw, I. 1999. *Qualitative evaluation*. London: Sage.

Smart, B. 1999. *Facing modernity: ambivalence, reflexivity and morality*. London: Sage Publications.

Sure Start Unit. 2000. *Sure start evaluation development project*. London: Department for Education and Employment.

Trigg, R. 2001. *Understanding social science: a philosophical introduction to the social sciences*, 2nd edn. Oxford: Blackwell.

Van der Knaap, P. 1995. 'Policy evaluation and learning: feedback, enlightenment or argumentation?' *Evaluation*, Vol. 1, No. 2, 189–216.

Walker, R. 2000. 'Welfare policy: tendering for evidence' in H.T.O. Davies, S.M. Nutley and P. Smith (eds), *What works? Evidence-based policy and practice in public services*. Bristol: Policy Press.

Weiss, C.H. 1995a. 'The haphazard connection: social science and public policy', *International Journal of Educational Research*, Vol. 23, No. 2, 137–50.

——. 1995b. 'Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families', in J.P. Connell, A.C. Kubisch, L.B. Schorr and C.H. Weiss (eds), *New approaches to evaluating community initiatives, Volume 1: concepts, methods and contexts*. Washington D.C: Aspen Institute.

White, M. 1999. 'Evaluating the effectiveness of welfare to work: learning from cross-national evidence', in C. Chitty and G. Elam (eds), *Evaluating welfare to work*. In-House Report No. 67, London: Department of Social Security.

Date received 12 February 2001. Date accepted 21 June 2001.