

Commissioning evaluation: is there an enlightened approach?

Michael Clegg

Introduction

This short paper is intended to be exploratory. It starts from the largely unevidenced premise (or hunch) that the art of commissioning evaluation has received little attention within the wider evaluation community, and that there has been limited reflection on practice by commissioners themselves. Certainly it seems there is a limited sense of common identity amongst those involved. The ideas put forward here will cover some familiar territory to many involved in evaluation. However, by discussing them in the context of commissioning I hope to raise wider questions of how reflection on commissioning might proceed and whether there is an appetite for building a more coherent identity amongst those involved. Comment is welcome from commissioners, evaluators and policy makers.

These remarks need two caveats. First, my recent experience comes from working at the New Opportunities Fund; that is outside the established communities of analytical services functions within Government Departments, or the larger frameworks of the Government Statistical Service and Government Social Research. Departments will have their own procedures, cultures and networks for those involved in commissioning. Whilst I reference some Government sponsored evaluation in the paper, I would generalise my observations to this sector only with caution. I would also emphasise that the views expressed here are my own, not those of the New Opportunities Fund.

Second, counter to the above assertion that commissioning has received little attention within the wider evaluation community, the UKES's own recently published draft good practice guidance for evaluation includes a section on *Guidelines for Commissioners*. This usefully lists the responsibilities of the commissioner to evaluators, applicant evaluators and to the subjects of evaluation, as well as their role in promoting the integrity and worth of evaluation. The Social Research Association has also recently completed a revision of its commissioning guidelines. However, this paper aims to address more general topics: what type of work do commissioners seek, does it match what evaluators might reasonably be expected to deliver, and what type of knowledge will be produced?

The paper poses the question, is there an enlightened approach to commissioning evaluation? Such attention to evaluation commissioning raises questions about the role of the commissioner. The DH *Research Governance Framework* usefully sets out the responsibilities of what it terms research sponsors (e.g. value for money and research quality.) However, whilst the *Framework* distinguishes between "Research Employer" and "Principal Investigator" no such distinction is made between commissioning organisation and commissioner. For evaluation at least, such a distinction is useful. Commissioners will represent the interests of their organisation to evaluators (ensuring value for money and quality, but also the policy relevance of work.) However, they also have a responsibility to communicate the interests of evaluators and evaluation to their organisation: protecting the integrity of individual pieces of work, ensuring findings are considered and disseminated, and informing policy makers as to what evaluation can reasonably be expected to deliver.

Use of the term *enlightened* in my title is intended to evoke Weiss's model of research utilisation – where the impact of research produced knowledge is indirect, creating a

context for policy making, not determining individual decisions. I hope to show both that commissioning can be usefully informed by reflection, and that good commissioning is about creating a context in which evaluation can inform policy.

Realistic evaluation?

In a speech to the UK Evaluation Society conference in 2000, reproduced as *Reflections on Evaluation in Practice* in *Evaluation*, vol. 7 no. 3, Janet Lewis noted an increasing tendency to pack evaluation specifications with major issues.

I do think we need to become better at defining the kind of evaluation it is realistic to carry out. I keep coming across suggestions for evaluations that are attempting to solve all the difficult questions of whether a whole range of activities have had the intended effects of changing the world, and whether doing something else would have been a more effective way of achieving these ends, and all on a modest budget. It is like suggesting that we should be doing synchronised swimming when all we can manage is rather bad crawl.

In the same volume, Robert Walker takes up a similar theme in relation to four of the Government's Welfare to Work pilot programmes. Walker notes that each pilot was characterised by multiple (sometimes imprecisely defined) policy objectives, with evaluators asked to assess the whole range. In each case evaluations centred on large-scale surveys in action and control areas, with the aim of assessing programme impacts. However, each also contained significant formative elements, generally involving qualitative interviews with clients staff and other stakeholders, looking at implementation and the effectiveness of partnership working in multi-agency delivery consortia.

Without immediately endorsing Walker's critique of its deliverability, we can see this omnibus approach as an emerging standard for evaluation commissions, well beyond welfare policy, or piloting. The common lineaments – an ambitious multi-method approach combining quantitative impact assessment with a concern for implementation and process, approached through qualitative work – can be seen in three brief examples.

- The specification for work on Sure Start, produced by the inter-Departmental Sure Start unit, required an “evaluation strategy with three distinct, but inter-dependent elements: implementation (or process or formative) evaluation; impact (or summative) evaluation; [and] an assessment of ... cost effectiveness.” The evaluators have responded with a multi-stranded study, prominent in which are (a) an assessment of impact on children, families and communities (through a longitudinal, quasi-experimental design using area controls and existing data from the Millennium Cohort Study;) and (b) examination of implementation, focussing on issues such as access and perceived quality of services, and with data collection including a repeated national survey of projects and 25 in-depth case studies.
- Work commissioned more recently by DfES, to evaluate the Connexions service (which provides information, advice and support for young people) has disaggregated the model: letting separate tenders for work on impact evaluation (looking at employment and employment related soft outcomes,) and on customer satisfaction and stakeholder perspectives (Bev Bishop, 2002) (It should be noted that in the case of the Sure Start and Connexions evaluations Lewis's strictures about a modest budget may not be applicable.)
- At the New Opportunities Fund we have drafted a generic evaluation framework to inform specifications for work on individual programmes. This contains three core elements: implementation evaluation (including construction of a typology of interventions within a programme, and assessment of the conformance of delivered

activities with programme aims, via case studies;) impact evaluation (generally including some kind of before-and-after design with a naturally occurring control,) and organisational evaluation (with a particular focus on delivery partnerships.) These are set in the context of understanding relevant theories of change.

For the commissioner this model offers a number of advantages. There are the widely accepted benefits of a multi-method design (Clarke, 1999: 89) with triangulation providing checks on the reliability and validity of data (and conclusions,) and an ability to interpret outcomes information through a knowledge of implementation process. More pragmatically, preliminary early findings supported by qualitative, implementation focused work are a boon for commissioners needing to satisfy their own organisations and stakeholders with early results. More pragmatically still, commissioners know that different types of data will be well received by different stakeholders: the killer statistic for policy makers, the quotation for lobby groups, and so on.

Both the theoretical and practical advantages noted are significant, and the examples quoted seem likely to produce good results. However, there is a need to be alert to the dangers of over extension flagged by Lewis, and which this model might encourage or disguise. Returning to Walker's critique of specific large-scale evaluations, this might be paraphrased as involving:

- an expectation from commissioners that the evaluation will address all the policy questions raised by a programme, when this is itself likely to have imprecise objectives and boundaries, a variety of implementations, and be subject to change.
- an emphasis on quantitative impact assessment through large-scale survey, challenged by the inherent difficulty of identifying a strong counter-factual (with this usually involving area controls.)
- complementary qualitative work which is in danger of becoming over-burdened with expectation (given weaknesses in impact assessment - see above - and the demand for early outputs from policy makers) but which cannot by its nature provide the clear, unambiguous answers desired.
- difficulty in integrating quantitative and qualitative elements of the study, with these often led from different institutions.
- a context where researchers are expected to deliver definitive evidence to support specific policy decisions.

The problem of an adequate counter-factual is one I am not equipped to address here. However, the broader points about over-extension and expectation have a general relevance. A first conclusion for enlightened commissioning is, therefore, that a useful model of multi-method evaluation has emerged, but that this can hide dangers of over-freighting any specific commission. Commissioners will need to pay attention to managing expectation within the commissioning organisation and its stakeholders. This theme is taken up with regard to accountability and attribution in the next section.

Surplus As: accountability and attribution

Discussions of evaluative activities and purposes uniformly include *accountability* as a primary benefit from evaluation (most often with some notion of *learning* as its contrasting pair.) In the article by Janet Lewis quoted earlier, Chemlinsky's three way distinction of types of evaluation is referenced, comprising evaluation for *accountability*, evaluation for *development* and evaluation for *knowledge*. Lewis notes that

the standard “worthy” definitions of the purposes of evaluation [of which this is one] are known to those of you who are practising evaluators, but they are not always familiar to people who are commissioning evaluations and therein can lie confusion.

She goes on to note that “evaluation for *accountability* is held in high regard [original emphasis]” and to envision as a counter-weight “a campaign for more evaluation for learning – of both process and impact”.

Despite this claim for its “high regard”, evaluation for accountability is at odds with much current practice. Evaluation findings are not linked to any framework of punishment or reward, as accountability implies (and if they were we would call them something different - inspection or performance management.) More fundamentally, the (dominant) realist paradigm in evaluation stresses that outcomes are the product of both programme mechanisms and each particular context in which the programme operates: the antithesis of the single judgement that accountability demands. Economic or audit traditions will stress accountability for the effectiveness (or economy or efficiency) of a programme in relation to original aims; but this seems to miss the point of a *what works* agenda where problems and failures are the basis for improvement. Indeed in the context of evidence-based policy and practice, an assessment of programme impacts against aims (nuanced for different contexts) can usefully be recast as a prerequisite for learning, rather than a means to accountability. Lewis usefully states the need for commissioners to be clear about the purpose of work they commission, but I suggest commissioners can go further and prioritise learning at the expense of accountability.

Attribution is often linked as an issue to accountability. If we wish to pass judgement on a programme it is necessary to define its particular contribution to changes in individuals or organisations (and this has become an acute issue given the large number of initiatives launched in recent years.) However, attribution also raises more fundamental (and interesting) questions of causation. Complexity theory (Sanderson, 2000) suggests that it would not be possible to work through strata of layered initiatives, parcelling out impact to each layer. Instead the interaction and feedback between the elements of a policy environment (including different programmes) mean that we should not expect, even with an effective programme, that the relationship between inputs (i.e. programmes) and outcomes will be linear. One programme may be a precondition for change, but that change will not be manifest until other elements are in place: change will not be incremental, but radical and swift.

Although this is very tentative, we have some indication of the working of complex systems with New Opportunities Fund programmes. The evaluation of our investment in computers in libraries, and related staff training, has given early indications of some fundamental changes in the way library services operate, who their users are, and so on. At this stage it seems plausible that the Fund’s programme has precipitated a change in state, made possible by changed expectations for library services as part of the Government’s broader social inclusion agenda. The other side of this coin is, of course, that some interventions may appear to have limited impact, but be necessary to later change. Commissioning organisations ignoring these facts of complex systems will inevitably find many evaluative findings disappointing.

Preliminary conclusions

The conclusions drawn above are challenging for commissioning organisations: there is a need for modesty about what a specific programme evaluation can deliver, there will not be a single outcome but varied, context sensitive outcomes, and definitive attribution of

change to the intervention will be difficult. What can the enlightened commissioner offer his or her organisation instead? Evaluations will need to make a contribution to our understanding of the mechanisms through which a programme produces observed effects in particular contexts, that is they will develop theory.

Theory

In suggesting an emphasis on building theory, this argument is following a fairly well trodden path (and theory here should be taken to mean domain specific, middle-range theory.) In the article from *Evaluation* already cited, Walker notes that whilst economic concepts, such as substitution and displacement effects, make some appearance in evaluation literature, reference to ideas in other social sciences is limited. He proposes a move away from assessment of specific programmes (and an expectation that unambiguous knowledge of what works will determine specific policy decisions in the short or medium term) towards review and studies developing theory. This is based on a model of how evidence impacts on policy making in which new knowledge accretes slowly – too slowly for the immediate policy making cycle –forming the broad structure for policy thinking: a similar model to Weiss' enlightenment.

Similarly Lewis fleshes out her proposed “campaign for more evaluation for learning” with a proposed “learning loop”. Theories of change should be both a tool for evaluators to pose questions about why particular interventions have particular impacts, and should be made explicit in the design of any intervention.

Whilst these proposals are persuasive, there is an absence of examples of what theory development in a policy area might look like. One useful example of which I am aware comes from work by Professor Nicholas Emler on self-esteem, published by JRF (and coming from within social psychology, this represents exactly the use of social science concepts Walker requests.)

Emler's work reviews existing research (evaluations per se are not the focus) on self-esteem, causes of high and low self-esteem, and its causal links to anti-social behaviours. He notes that whilst low self-esteem is a core element of popular theories explaining behavioural and social problems (including the theoretical stance of those making policy) the most reliable evidence points to very different conclusions. Specifically, relatively low self-esteem is not a risk factor for delinquency, violence towards others, drug or alcohol use, educational under-achievement or racism; young people with relatively high self-esteem are more likely than others to hold racist attitudes, reject social pressures from peers and parents and engage in physically risky pursuits such as drink driving. To reiterate, whilst Emler's work is not primarily evaluative, it shows what can be achieved through attention to developing theory: a challenge to predominant folk theories; suggestion for the direction of future policy; and the definition of areas where empirical findings need more work to secure full understanding (one area of the work that does look at the evaluation of programmes concludes that interventions to raise self-esteem can have an impact, but knowledge of why particular interventions work is limited.) It is also worth noting that Emler's work illustrates how policy-oriented work is one field where a null result can provoke a high degree of interest.

However there are also more practical issues. As discussed earlier, the commissioner operates somewhere between the evaluation practitioner and the decision makers of his or her own organisation and its key stakeholders. For many outside evaluation commissioning or practice, including those broadly sympathetic to evaluation and

evidence-based policy making, “theory” is a trigger word for antipathy. No amount of explanation that, for example, *theories of change* means a careful understanding of their own ideas, and those of front-line delivery staff, will overcome hostility generated by the phrase. Lewis’ use of *learning* is helpful here, but evaluation practitioners will need to help commissioners develop new ways of presenting these ideas if decision makers are to be brought on board. Further, evaluators themselves will need to respond to this agenda, it is a two way process: evaluative questions should be informed by existing theory; standard instruments and scales should be used where possible; and there would be great value in evaluators sharing knowledge about intermediate indicators. More broadly (and taking the need for robust conclusions as a given) there needs to be a degree of boldness in the move from empirical findings to theoretical generalisations.

Conclusion: characteristics of enlightened commissioning

Whilst good evaluative work – and more ambitiously a thriving evaluation sector – needs good practitioners, it also needs intelligent and reflective commissioning. I suggest that key issues for enlightened commissioning will be:

- modesty in expectation, reflected back to the commissioning organisation;
- a focus away from accountability and attribution;
- an orientation towards developing theory;
- a concern to make a reality of multi-method approaches, so there is genuine triangulation and mutual illumination;
- an openness to complexity and systems approaches.

References

- Bishop, B. (2002) “The evaluation strategy for the Connexions service”, *2002 DfES Research Conference Papers*
- Clarke, A. (1999) *Evaluation Research* London: Sage
- Davies, H., Nutley, S. and Smith, P (eds) (2000) *What Works? Evidence-based policy and practice in public services* Bristol: Policy Press.
- Department of Health (2001) *Research Governance Framework* London: DH (www.doh.gov.uk/research/rd3/nhsrandd/researchgovernance.htm)
- Emler, N. (2001) *Self-esteem: the costs and causes of low self-worth* York: JRF
- Lewis, J. (2001) “Reflections on evaluation in practice”, *Evaluation* 7(3)
- Sanderson, I (2000) “Evaluation in complex policy systems” *Evaluation* 6(4)
- Sure Start Unit (2000) *National Evaluation of the Sure Start Programme in England: Specification of requirements* London: Sure Start (www.surestart.gov.uk/infoeval)
- Walker, R. (2001) “Great Expectations: can social science evaluate *new labour’s* policies?”, *Evaluation* 7(3)
- Weiss, C.H. (1979) “The many meanings of research utilisation”, *Public Administration Review*, 39 (5).

Information on New Opportunities Fund evaluation is available at www.nof.org.uk/index.cfm?loc=gen&inc=research/index