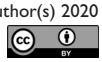
Article

Evaluation
2021, Vol. 27(1) 32–56



DOI: 10.1177/1356389020976157
journals.sagepub.com/home/evi


How does the commissioning process hinder the uptake of complexity-appropriate evaluation?

Jayne Cox
Brook Lyndhurst, UK

Pete Barbrook-Johnson
University of Surrey, UK

Abstract

This paper investigates the role of evaluation commissioning in hindering the take-up of complexity-appropriate evaluation methods, using findings from interviews with 19 UK evaluation commissioners and contractors. We find, against a backdrop of a need to 'do more with less' and frustration with some traditional approaches, the commissioning process is perceived to hinder adoption of complexity-appropriate methods because of its inherent lack of time and flexibility, and assessment processes which struggle to compare methods fairly. Participants suggested a range of ways forward, including more scoping and dialogue in commissioning processes, more accommodation of uncertainty, fostering of demand from policy users, more robust business cases, and more radical overhauls of the commissioning process. Findings also emphasised the need to understand how the commissioning process interacts with the wider policy making environment and evidence culture, and how this manifests itself in different attitudes to risk in commissioning from different actors.

Keywords

behaviours, commissioning, complexity, evaluation, policy

Introduction

Interest in using ideas and methods from complexity science to enhance policy evaluation has increased in the last decade or so (Walton, 2014, 2016). A perceived benefit of 'complexity-appropriate' evaluation is its ability to capture the full complexity of the policy and context

Corresponding author:
Pete Barbrook-Johnson, Department of Sociology, University of Surrey, Guildford, Surrey GU2 7XH, UK.
Email: p.barbrook-johnson@surrey.ac.uk

being evaluated (e.g. path dependency, emergence, feedback loops, multi-causality; see Boehnert et al., 2018), using methods that can adapt to emerging findings, that involve iteration and multi-stakeholder working (Barnes et al., 2003; Gates, 2016; Mowles, 2014; Reynolds et al., 2016; Sanderson, 2000; Walton, 2014; Williams, 2015).

Competitive tendering through government research frameworks is an important way in which policy evaluation is commissioned in the United Kingdom. The aim of this paper is to identify barriers to the uptake of complexity-appropriate methods that arise from the commissioning process, and practical changes to make commissioning easier if barriers were found. We report the research findings from qualitative interviews with 19 commissioners and contractors who have been involved in tender exercises for policy evaluation studies, mainly for one UK government department.

The paper begins with a summary of relevant work on evaluation commissioning, then describes the scope and method. The findings cover the wider operating context for commissioners and contractors as an influence on their attitudes and behaviours during commissioning, then the perceived barriers within the commissioning process, first for novel methods in general, then for complexity-appropriate approaches and methods specifically. We then describe interviewees' suggestions on practical measures for enhancing the uptake of complexity-appropriate evaluation and conclude with the authors' reflections.

Relevant work: Recent practical guidance for commissioning in complexity addressed a gap in the academic literature

In this section, we first review relevant academic research which has tended to only consider commissioning in passing or as part of a wider study, or to propose changes which make sense for evaluators, but not for commissioners or users. We then review practitioner guidance and discuss why these, despite only offering partial resources, are so important in the political environment evaluation operates in.

Academic research

There is a relatively small selection of studies which have dealt directly with the commissioning process and its impact on evaluation. Where it is explored, it is typically part of a wider study on evaluation or applied academic research as a whole. The differences between commissioners and evaluators is a common theme; Schneider et al. (2016) and Broer et al. (2017) find that evaluators and commissioners sometimes differ in what they perceive as 'good' evaluation and typically have different views and priorities. Gates (2017) suggests systems and complexity approaches can be a way to bridge this gap and be used to reconfigure the relationship between commissioners and evaluators, away from a client-service-type relationship, towards partnership models involving shared question setting, negotiated and flexible contracts, and a high level of involvement of commissioners in the actual evaluation activities.

Others have identified barriers (within and around commissioning) which can undermine effective evaluation; Schneider et al. (2016) finds a long list of barriers which the commissioning process creates (note, they do not focus on complexity or systems-appropriate evaluation specifically), including political influence, tight funding, unhelpful timeframes, lack of a 'culture of evaluation', caution over anticipated results, and lack of skills among commissioners. The LSE GV314 Group (2014: 226) are especially concerned with political influence,

including *"political ammunition objectives' [i.e. wanting to use evaluations to support already held political positions] as opposed to 'scientific relevance and quality objectives"* when designing an evaluation, which they propose can shape the tender specification and compromise methods choices. Walton (2016) takes a step back and suggests it is the framing (i.e. rhetoric and political context) and governance of a policy or programme itself which may create barriers for effective complexity-appropriate evaluation, such as too narrow focus, or a limited set of evaluation users and stakeholders. Barbrook-Johnson et al. (2020) find mixed understandings and different views on the value of complexity as a lens for evaluation, among commissioners, suggesting negative views or misunderstandings (e.g. believing that being complexity-appropriate is more expensive, or creates room for uncertainty to be used as a defence by evaluators for lack of rigour, precision, or clear recommendations) may be a barrier in some cases. Barbrook-Johnson et al. (2019) more fundamentally suggest profound inflexibility, conventionality, and inertia in the research and evaluation commissioning process undermines potential for complexity-appropriate evaluation. None of these studies pay explicit attention to the rationale for the features of the commissioning processes that create the barriers and how these confer benefits to the organisations commissioning the evaluation (e.g. internal clients want them, they feel secure against critique).

However, De Laat (2013) do consider the relationship between evaluator, evaluand and commissioner in depth, but focus on the role of the commissioner (which they rightly highlight is largely omitted in the evaluation literature). They do not focus on the commissioning process in detail, but rather propose a framework through which to understand the role of commissioner. Nonetheless, this is a valuable contribution in light of the focus of existing studies on suggesting changes to the commissioning process and those which view things solely from an evaluator standpoint.

Proposing changes to the commissioning process

Another theme, which we pick up in this paper too, relates to how the commissioning process might be improved or changed. Gates (2017) makes some proposals to this end, in the face of systems and complexity thinking, suggesting different 'things' should be evaluated (i.e. more complex or systemic issues, boundaries of evaluations should be re-drawn), and the process should be designed to be more flexible and adaptive. Giorgi (2017) digs further into practicalities, finding that commissioning is often overly separated from other components of the evaluation and policy making process, despite efforts and the intention to have them be more integrated. Similarly, Anderson et al. (2008) suggest commissioners need to be more explicit about the aims and intended uses of studies (presumably, where they are vague in tenders, or where there are other 'real' or political intentions, beyond those included in a tender), and that they need to have regular contact with evaluators, though this is common in many places during an evaluation, but not during the commissioning process. De Laat and Williams (2013) pull out lessons for evaluation commissioners based on experience of evaluations in the European Commission, though they do not focus on complexity specifically. Lessons for the commissioner include, providing results on time, using steering groups, continuous quality assurance, developing dissemination strategies, and being acquainted with wider budgetary and policy processes.

These lists of things to improve or change are helpful, but only take us so far. In this paper, we build on these and go further, to develop a richer understanding of how these requirements

interact with the hard reality of commissioning processes, the wider policy process, with different epistemological views, and demanding policy users. Though not focussed on evaluation, but rather on the commissioning of public service delivery in the face of complexity, Davidson Knight et al. (2017) demonstrate one way to do this. They describe an emerging approach to commissioning public services in complex environments, emphasising an optimistic message that there is an alternative to traditional approaches, which break down the conventional arms-length contractor–client relationship. Our aim here is pragmatic: to describe what commissioners and contractors believe can be done within, or beyond, the existing public procurement system to enable complexity-appropriate evaluation. We do not intend to define or reflect on the value of complexity-appropriate evaluations, which many others, including those cited in the introduction, have done.

Patchy and evaluator-focused

The studies with some direct focus on the commissioning process in evaluation are limited both in the focus they put on commissioning specifically (many are focussed on a wider question), but also in their geographic and policy domain coverage, which tends to be dominated by experiences in Australia and the health domain, respectively. Relatedly, Schneider et al. (2016: 209) also note, *'[t]here are also few studies that deliberately seek to gather, compare and analyse perceived barriers from both policy makers and evaluation researchers operating within the same policy space'*. There appears an implicit bias in much of the literature, taking the view of evaluators alone, rather than engaging more with the pragmatic reasons on the policy side that suit things being as they are. Other excellent studies omit commissioning as a significant concern or only refer to it implicitly, for example, Stame (2010) explores why evaluations fail, asserting they look beyond methodological concerns, but does not focus on commissioning as component in and of itself. Similarly, Walton (2016), Reynolds et al. (2016), McIntosh et al. (2018), and Schwarzman et al. (2018, 2019) all consider conceptual and practical barriers to the wider adoption of (complexity-appropriate) evaluation, but either do not tackle the commissioning process directly, or mention it only in passing. Why this is, is unclear.

Practical guidance

Until recently, there were comparable gaps in the practice literature, where many different organisations produce their own guidance for evaluators. However, these have been partially addressed by the recent publication of the updated UK Magenta Book and its Annex 'Handling Complexity in Policy Evaluation' (HM Treasury, 2020), and a Complexity Evaluation Framework, or 'CEF', developed for its own commissioners by one UK government department (Defra, 2020) but available to all. The Magenta Book annex on complexity overviews why complexity matters in evaluation, the challenges it poses, makes suggestions for commissioning and managing in complexity, and selecting approaches and methods. On commissioning, it articulates the need to rethink what evaluation is and to adjust expectations in the context of complexity, both to be more ambitious in how evaluation connects to the rest of the policy process, but also to be more realistic and appropriate in demands put on evaluations for quantitative rigour. As a high-level piece of guidance, the annex does not dive into the details and politics of the commissioning process, but rather provides a set of heuristics or general-purpose

pointers on how to commission in complex settings. These include challenging tradition notions of evaluation and evaluation design, connecting tightly to other parts of the policy process, engaging with stakeholders, and making evaluation management flexible.

The CEF developed by Defra is an example of how the high-level guidance in the Magenta Annex has been tailored to the specifics of one department. Its guidance provides a set of questions as prompts (and related suggestions) to cover a wide range of technical issues when evaluating in complex settings. Like the Magenta Annex, though, this guidance leaves negotiating the actual day-to-day details of the commissioning process to readers. This is entirely understandable given the purpose and scope of both new sets of guidance but does mean that despite the welcome addition of both, commissioners and evaluators across government are still left without detailed guidance on the political, legal, financial, and procedural day-to-day realities of commissioning in complex settings.

Beyond these recent additions, most guidance does not mention complexity at all, or only in passing; others identify methods that could be used in complexity-appropriate evaluation (e.g. Stern, 2015, for development evaluators) and some practitioner-focused websites offer information or training opportunities on complexity-focused evaluation (e.g. UK Evaluation Society, betterevaluation.org). A lack of coverage and detail on complexity and commissioning in official government guidance (which is only partially addressed by the recent annex and CEF) is relevant here because commissioners (in the United Kingdom at least) are expected to adhere to the approaches and methods set out in the guidance. Individuals are further bound by formal quality assurance responsibilities to uphold evidence standards (e.g. the Government Social Research Code, Government Social Research Profession, 2018). While not relating to complexity specifically, recent wider debates on formal evidence standards (see Puttick, 2018, for example) and evidence-based policy in government are therefore also relevant. For example, an enquiry by the National Audit Office (2013) concluded that government should review commissioning arrangements with a view to adopting higher quality and more robust – by which it meant quantitative and experimental – approaches in policy evaluation. Davoudi et al. (2015) discussed whether such a drive to common evidence standards might result in too narrow a focus on certain approaches, squeezing out other methods that do not fit the narrow definition of evidence quality. Also concerned with evidence-based policy, Head (2010) mentions the role of bargaining, entrenched communities and multiple stakeholder interests and values as limitations on rational decision-making. Reporting on policy stakeholder discussions, Rutter (2013) similarly identified barriers to the wider use of evaluation evidence that arise from cultures and incentives on the demand side (i.e. policy clients), including avoidance of political risk. How such tensions might play out in evidence commissioning is not covered.

Attempting to partially address these gaps is our aim here, by focussing on the commissioning and tendering process specifically, by focussing on environmental policy domains in the United Kingdom, and by speaking to individuals on both sides of the supposed commissioner-evaluator divide.

Scope and method

We theorise that the approaches and methods selected in commissioned evaluations evolve from the behaviour of actors on both ‘sides’ of competitive tendering, interacting with a rules-based procurement process (e.g. derived from public procurement regulations).[1] Behaviours

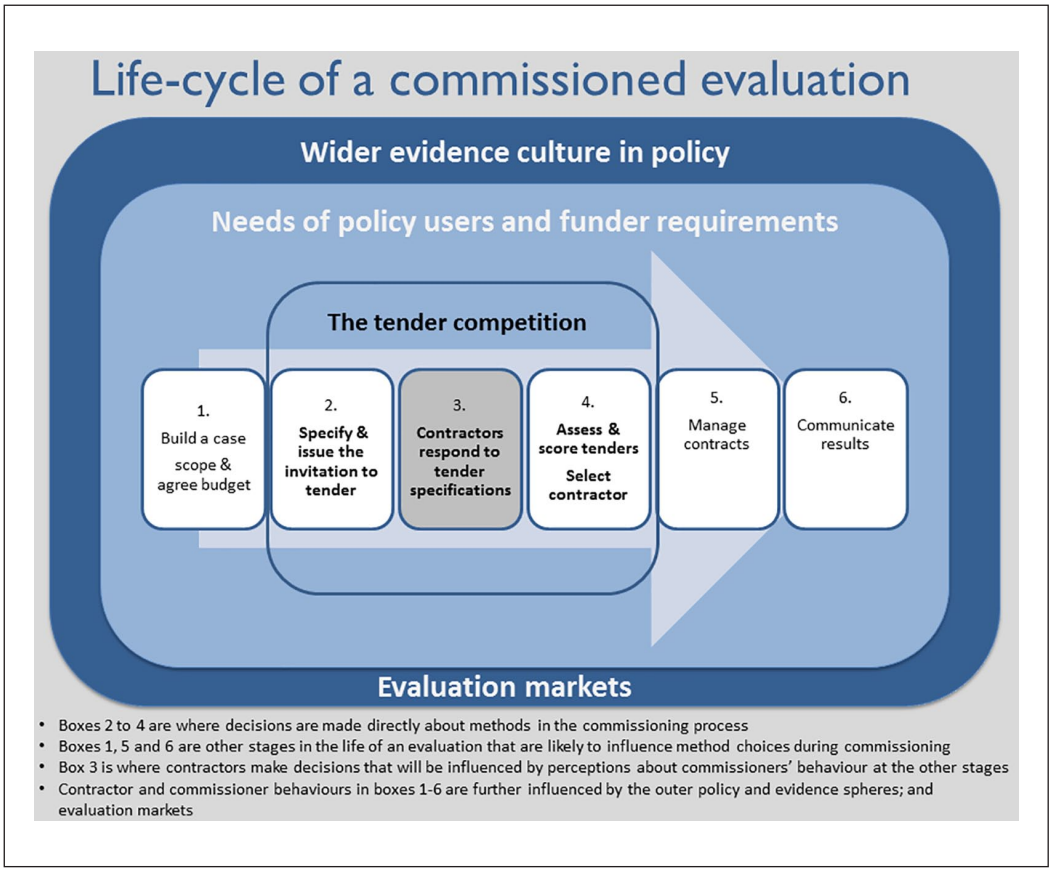


Figure 1. Simplified description of the different stages in the lifecycle of an externally commissioned evaluation.

are further influenced and shaped by the contexts in which the actors are operating, including the political and intellectual zeitgeist. The tender competition is the decision-making nexus where rules and behaviours interact and where the wider influences on commissioners and contractors are crystallised in the choices that are eventually made about approaches and methods.[2] Sources of influence throughout the lifecycle of a commissioned evaluation are illustrated in Figure 1.

A semi-structured interview guide was developed around the framework in Figure 1, to explore the immediate constraints from commissioning on decision-making with respect to methods, as well as 'upstream' and 'downstream' influences. The main question themes and flow are summarised in Figure 2. To aid discussion, basic concepts were introduced to those who were not familiar with complexity, focusing on aspects such as emergence, feed-back loops, multiple causal paths and so on. The research took place before the updated UK Magenta Book and its Annex 'Handling Complexity in Policy Evaluation' (HM Treasury, 2020) was published, so this was not covered.

In-depth interviews were conducted in summer 2018 with 9 commissioners and 10 contractors who are directly involved in evaluation tendering processes. Figure 3 describes the full

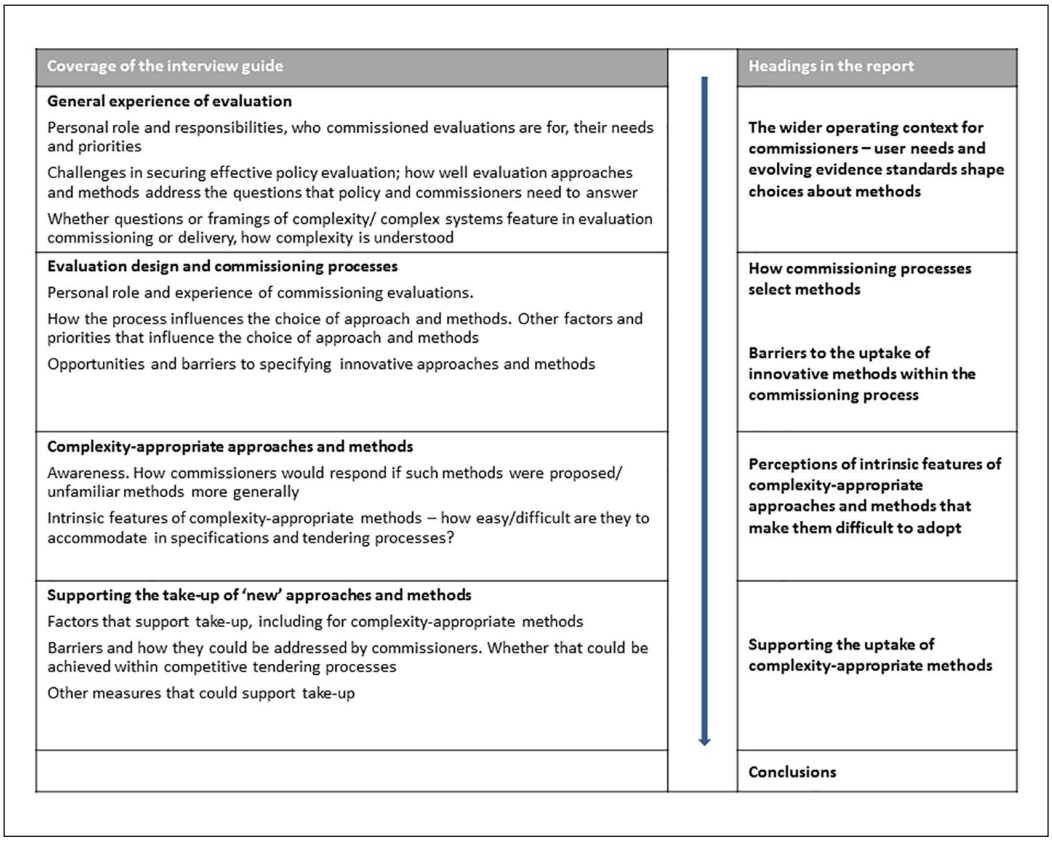


Figure 2. Summary of themes in the interview guide.

methodology, including the sampling approach, data extraction and management, and analysis. A thematic approach was used for the analysis. The manual procedure (feasible because of the small sample size) involved several steps and iterations. Data for each case (interview) were summarised systematically from transcripts in a framework grid developed in MS Excel, where category headings were derived from topic guide themes and key questions, plus a few additions from an initial coding of transcripts. The categories included descriptive, contextual information (e.g. contractor or commissioner, evaluation role and expertise) to enable analysis by different respondent characteristics and contexts, and for the whole sample. Emergent themes relating to the categories were derived from initial coding of transcripts and during detailed summarising. Further themes and connections were identified through both cross-case and within-case searching and mapping. Findings were 'tested' at two sessions with evaluators in government and contractors, and academics, which prompted clarification of some of the findings but no major revisions.

Coverage of the interviews was limited to one department to ensure consistency in the contexts and procurement processes being explored. That does not mean, however, that the commissioning process described in the paper is unique to the department. Evaluations may be commissioned through pan-government procurement frameworks and several interviewees gave examples of similar practices and barriers in evaluations commissioned in other departments.

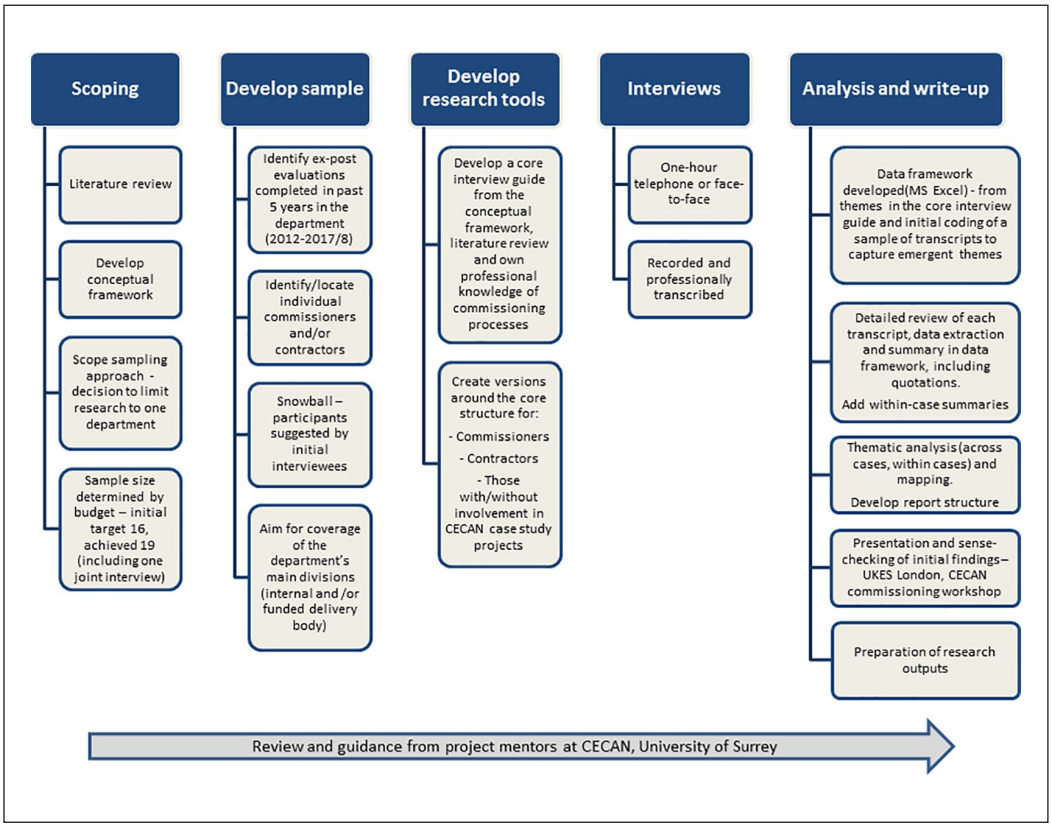


Figure 3. Summary of the methodology.

'Commissioners' were social research managers and analysts, mainly in one UK government department and its network (i.e. bodies who support policy delivery). They are directly involved in specifying evaluations, preparing tenders and managing contractors, often as part of a wider research/evaluation role. They are typically accountable to internal policy 'clients' and funders for the scope and outputs of evaluations. 'Contractors' were senior individuals in private companies who devise and lead evaluation bids. The sample was identified by first identifying evaluations completed for the department in the past 5 years, then by cascading from contacts known to the authors or from initial interviewees. It is important to note, individuals gave their personal views and the findings should not be taken to represent the position of any individual, public body or private organisation.

While interviews were spread across different policy areas there are limitations from the size and nature of the sample: (1) we could not cover the entire commissioning chain in the policy areas covered, notably procurement officials or higher-level budget holders (e.g. deputy directors in the department), who might have different perspectives. (2) While findings were sense-checked in workshops with a wider policy evaluation audience, findings may not be repeated in other departments or beyond the United Kingdom, where procurement practices may differ. (3) While the contractors we interviewed work across UK government departments, there is a larger market of evaluation contractors that were not interviewed. Further research to test these findings more widely would therefore be worthwhile.

Findings

Throughout the findings we refer to 'commissioners' or 'contractors' where results were specific to that group. Otherwise, themes are drawn from contractors and commissioners alike. Findings broadly follow the order of themes in the interview guide: Figure 2 shows how the headings below refer to interview themes. Behaviours around risk emerged as an important cross-cutting theme.

The wider operating context for commissioners – User needs and evolving evidence standards shape choices about methods

Commissioners and contractors discussed how the wider context in which evaluations are developed and used bears on the decisions they make when specifying or responding to an evaluation tender. Considerations included

- **How results would be received by their 'clients' (policy users and funders)**, whether it would *'land'* with policy (be understood and meaningful); whether it would be seen as value for money, especially in a climate of austerity and *'doing more with less'* in government spending; and for some, whether those factors would create a risk to their own professional credibility or satisfaction.
- **A push to raise evaluation standards in a climate of austerity**. 'Landing' evaluation findings was also commonly discussed in the context of accepted evidence standards. Many welcomed a push to raise standards across government, in response to past criticism of weak practice (e.g. National Audit Office, 2013), including measures to raise the level of evaluation expertise within the department. At the same time, there was concern that raising standards had narrowed the range of methods that would be acceptable to policy evaluation funders – namely quantitative, counterfactual, and experimental approaches, including randomised control trials (RCTs). Quantifying impact was seen as the top evaluation priority. Several felt these methods now delimited the evidence 'quality standard' for evaluation, while case study and other qualitative approaches had fallen completely out of vogue.
- **A perceived lack of flexibility in budget processes that could discourage the trying-out of new[3] methods**. Commissioners especially criticised evaluation commissioning for its lengthy approvals process, often combined with the pressure to deliver evaluations within a single financial year for which the budget is available, and the length of time then needed to run a procurement exercise. Some commissioners felt this locked them into sub-optimal evaluation designs where policy had changed in the gap between budget approval and commissioning, and there was insufficient time to re-negotiate scope and go through the whole approvals process again. For both commissioners and contractors, it might also constrain methods choices to *'what can be delivered within the timescale and budget'*, which could rule out complicated data collection or methodologies that are designed for emergent causation (i.e. that which arises from the interaction of multiple actors in the system and is difficult, or impossible, to predict beforehand), for example. While this practice was reportedly common, a few felt that useful commissioning lessons could be learned from large budget, multi-year evaluations in high-profile policy areas that have used alternatives to conventional approaches, such as theory based and realist evaluation.

- **Differing levels of evaluation competency across policy areas influencing commissioners' readiness to explore less-familiar approaches**. Some commissioners felt they needed to embed the basic evaluation competencies and accepted methods first: one said that complexity-appropriate methods were *'a step too far'* for their colleagues. Others were more willing to consider alternative approaches and methods but still felt that the dominant evidence culture would constrain what could be commissioned. For some, this arose from an asymmetry in decision-making authority between evaluation funders/policy and social research managers when evaluation tenders are being developed.
- ***'Bolted-on'* evaluation limiting the range of methods that are feasible**. Commissioners and contractors both expressed frustration about evaluations where the approach was developed *after* the policy design had been fixed – or was even devised after delivery had started. This was reportedly common. They felt it left them with limited options for methods, with choices being driven pragmatically by data availability within the timescale and budget, rather than the best-fit approach for the evaluation questions.
- **Policy isn't asking for complexity-appropriate evaluation**. While many said their evaluations were 'complex', they often appeared to mean 'complicated'. Many described features of complexity that *manifested* in programmes and policies which they had evaluated (shown in Figure 4) but few were using complex systems as an *organising concept* for the evaluation framework and questions. The apparent mismatch between the design of delivery models and the linear framing of questions of impact was a challenge for delivering effective evaluation of some programmes. As a result, some felt that counterfactual approaches are not well-suited to the examples they gave but had not seen widespread demand from policy clients for complexity-informed or 'non-standard' (in their view) evaluation approaches (e.g. realist evaluation). Moreover, interviewees' knowledge of complexity-informed thinking differed: only half were familiar with complexity-appropriate evaluation, most often theory-based and realist approaches, and methods such as qualitative comparative analysis (QCA) and process tracing.
- **Complexity-appropriate evaluation challenges existing ways of thinking and doing**. Some who were aware of complexity science felt that shifting mind-sets from thinking about policy in linear to non-linear ways would be a difficult challenge.[4] That would include persuading policy of the value of asking different kinds of evaluation question. A few feared that complexity approaches might be seen as too '*academic*' to be useful for impact/accountability purposes or to improve policy: '*complexity is no different in that you need to make sure that the outputs are genuinely useful, otherwise it is a piece of academic work that doesn't land in practice*'.

The themes above emerged as important influences on how far commissioners and contractors are willing to depart from well-known and accepted (or *'standard*') evaluation methods, principally those set out in the UK Treasury guidance. In the prevailing budget and evidence climate, many felt it was difficult to take risks on 'new' or possibly expensive complexity-appropriate methods. They often preferred conservative methods choices as a result. To improve the appropriateness of evaluation, several (including contractors) would like to see a more plural evidence culture, to include a pragmatic blend of approaches from different evidence traditions.

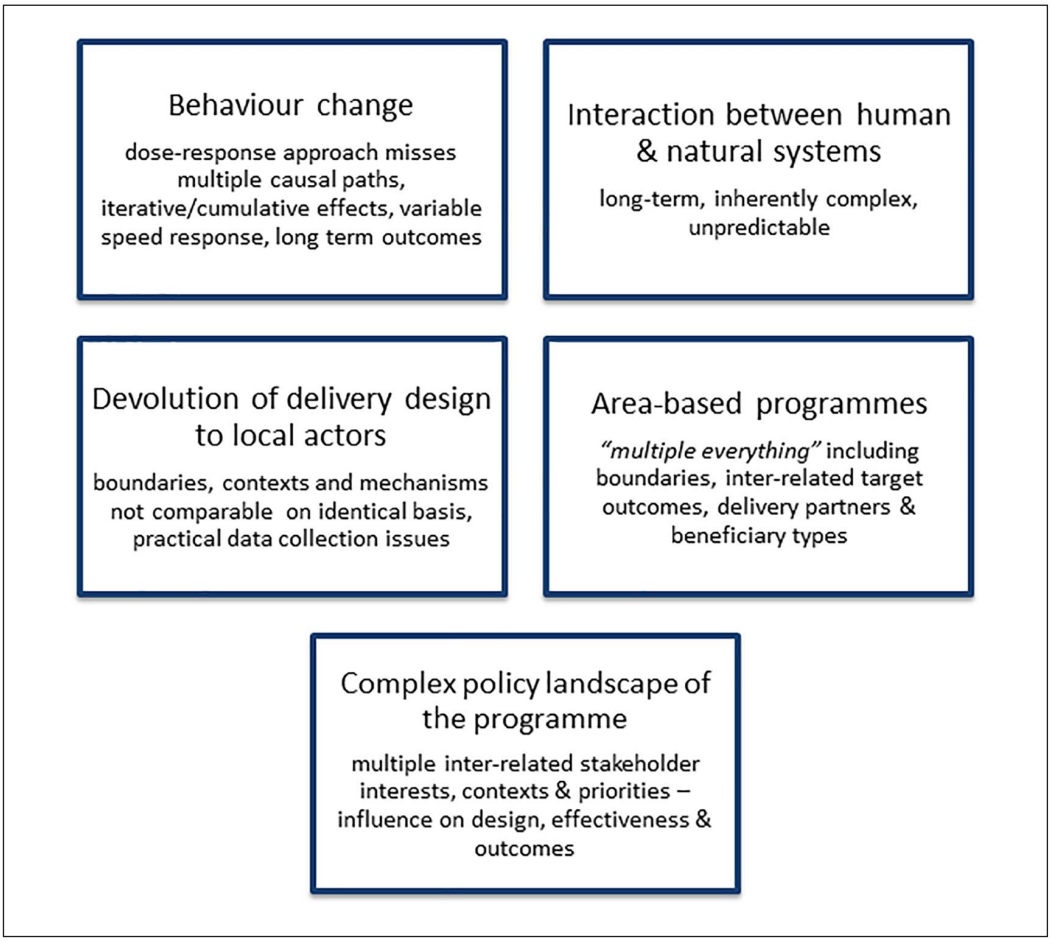


Figure 4. Manifestations of complexity where commissioners felt conventional evaluation methods had shown significant limitations.

How commissioning processes select methods

The interviews revealed how individuals' behaviours and decision-making processes interact with the formal commissioning structures to select the methods that are eventually used in evaluations. Here we explore how barriers to 'new' methods arise in commissioning.

The commissioning process – Multiple interests and priorities. The commissioning process described by interviewees typically involves individuals from (at least) three government 'professions', each having different priorities and interests in the evaluation. Normally they represent:

- Social science – typically the manager for the evaluation who is a specialist in research methods, though not necessarily evaluation.
- Policy – typically the policy or programme 'owner' and user of the evaluation findings, whose priorities are timely delivery and usability of the findings, subject to assurance about evidence quality.

- A procurement representative – concerned with ensuring a commercially fair competition within the relevant regulations, assuring the financial viability of contractors, and value for money of the proposed approaches and outputs.

To ensure a fair competition, contractors' bids will be scored by the assessors (typically the above individuals) against an assessment framework agreed with Procurement officials during the tender specification process. Scores for different aspects of the tender proposal will be given different weights. Each tender will be given an overall score, typically combined from: a technical quality score (methodology, appropriate outputs and contractor experience), a project management score (e.g. including risk management), and a financial score.

Crucially, the winning methodology will be the one that has scored highly across the various criteria, including lowest price. As we shall see later, the weighting between technical and financial scores is an important influence on contractors' decision-making about methods.

The choice of approach and methods in evaluation tendering processes – Balancing multiple requirements and appetites for risk. Decisions are made about methods at several stages during tendering: when formulating the tender specification, in the contractor's response, and in the tender assessment process to select a winning contractor. Interviewees reported that many compromises are made along the way and the final choice is rarely confined to purely the technical merits of a method.

Balancing multiple requirements. Commissioners set out their preferred methodology in the tender specification, which contractors then respond to in competition with each other. The final specification emerges from a negotiation between the social science lead, policy and procurement. Commissioners said it will reflect a compromise between the different priorities of those involved, which is often compounded by the short timescales for developing specifications. Lack of time may focus attention on well-understood methods. More rarely, commissioners had been able to undertake a '*scoping study*', either a full evaluability assessment (rare) or exploration of feasible evaluation methods for an extant programme. Aspects to balance when choosing methods were said to be:

- Policy clients' and funders' preferences (including external funders – e.g. if they are a department-funded body or the Treasury is involved)
- Appropriateness to the evaluation questions;
- The approach that was agreed with the funder;
- Combined constraints of budget and timescale;
- The limits of personal knowledge of methods;
- Government evaluation guidance (the Magenta and Green Books);
- Personal confidence in securing a useful and defendable result;
- Confidence that methods can be scored fairly and effectively in the tender assessment process so that the tender exercise will succeed.

A priority for commissioners is to issue a specification that will attract enough bids to ensure an effective competition. If a tender exercise fails, the lengthy approvals process may preclude a re-tender with a revised specification. Contractors will not always respond if they

feel an evaluation is unclear or unrealistic (e.g. scope, timescale, budget), too risky, poor quality, or does not add to their knowledge or professional standing. Principal influences on methods choices by contractors are a need to be price competitive, to offer added value over competitors (e.g. by being more creative or innovative), while remaining compliant with the commissioner's specification (e.g. meeting the tender scoring criteria).

Appetite for risk. Perceptions of risk emerged as one of the most widespread and important influences on the choices that are made about designs and methods during evaluation commissioning, in terms of what commissioners specify and what contractors propose. This included risks arising from the wider context described above (notably pressure on research budgets and meeting client expectations) and perceived risks of failure in the procurement process.

Commissioners, especially, appeared to be risk averse. They were more likely to stick with methods that had worked in the past, that they knew could be achieved within a given budget and would deliver an answer they were able to communicate (and defend) to policy. A further barrier, though less often mentioned, was lacking knowledge about the typical cost of an unfamiliar method to support a judgement about value for money. A few were concerned about their own lack of familiarity with non-standard evaluation methods, which could impair their competence to manage the contractor and quality assure the evaluation outputs.

More generally, uncertainty about what 'new' methods would deliver was a potential barrier, including whether that could be explained or assessed adequately in tenders. Some also flagged that it would be difficult to assess delivery risks adequately, whether that was risk of evaluation failure or resource risk from 'learning on the job' (with consensus that this investment risk is borne largely by contractors). For contractors specifically, risk also arises from uncertainty about future client demand to balance against the scale of investment that would be needed to upskill their teams in 'innovative' methods.

A few examples were given where risks of the unknown were mitigated through an element of risk sharing with the government evaluation client. For example

- Enabling flexibility to re-profile resources and deliverables at key stages during a large and complicated evaluation, or when using a novel method.
- Where the client and contractor have a long-term working relationship, which was said to enable shared learning over several projects and build mutual trust in finding solutions to cope with the unexpected.

However, those who spoke about it felt that procurement processes specifically discourage this type of co-productive learning relationship, with shared goals and risks, favouring instead an arms-length purchase-fulfilment relationship to avoid creating financial conflicts of interest.

Barriers to the uptake of innovative methods within the commissioning process

Procurement – Too narrow and rigid? While most interviewees make the best of the commissioning system, some questioned whether it is fully fit-for-purpose for evaluation and research commissioning. There were complaints that it prevents evaluators from responding either to evolving (*'agile')* policy making or the emergent conditions where policies are being implemented and evaluated.

Research procurement in government (including evaluation) has developed from systems designed for procuring a defined physical product, where suppliers compete on price to fulfil the 'delivery' of a specified number of items at a minimum quality or better. All parties in the competitive process need to be kept separate to prevent collusion and thus ensure that the procurement authority achieves an economically efficient price and best value for money. Similar principles apply to research and evaluation procurement: contractors bid against an overall requirement and list of 'deliverables' or outputs, pre-competition dialogue (e.g. to scope the feasibility and applicability of approaches and methods) is normally ruled out, and contractor performance is tied to the deliverables in the tender specification (which then forms the basis for the contract), with limited scope for variation during delivery of the contract. In this system, the tender specification plays a significant role in determining which methods are selected, as well as the scope for innovation.

Tender specifications – A guessing game for contractors that can favour conservative choices. Commissioners can signal their appetite for innovation through the tender specification, although both they and contractors described the process as imperfect for achieving that. The two options open to them are to issue: (1) a tightly defined specification, citing a specific approach, method and narrowly defined outputs; or (2) to offer a more open specification, where required outcomes are clearly defined but contractors have scope to interpret the best ways to achieve those.

Procurement officials (and some policy clients) were believed to favour tight specifications to '*create a level playing field*' and to ensure a successful competition. Tight specifications were thought to be useful where time and budget are tight, and where evaluation questions are simple. On the other hand, there was a view that this route can favour methods that commissioners already know well and therefore deters contractors from proposing new ones, because those would be less likely to '*tick the boxes*' in tender assessment frameworks. Some felt that tight specifications tend to use scoring criteria that favour tangible outputs, which can make it difficult for methods with less easily countable 'deliverables' to compete on an equal footing. One commissioner said that can favour contractors who '*promise the moon*' (i.e. large quantities for low cost); a contractor made a similar point about methods such as stakeholder workshops where cost per unit could be a poor indicator of quality and value for money (but, they say, is often used).

There was consensus that open, outcome-focused specifications, that invite 'creative' approaches, tend to leave more room for new methods to emerge, unless commissioners have the knowledge and skill to propose them in a tight specification. However, there are drawbacks that need to be managed. First, invitations for creativity may elicit too many radically different proposals which cannot be assessed effectively through the scoring framework and therefore risks a failed procurement (several described this as '*the apples and pears*' problem). There might also be disagreement about the scoring framework between the research manager and procurement: a few, for example, felt that procurement officials do not understand the implications and value to evaluation outcomes of different methods choices. Second, contractors have to second-guess what commissioners really want and mean by 'creative', which may result in contractors being cautious in what they propose, to ensure they stay compliant and price competitive.

Commissioners can provide clues – for example, the weighting for price, the overall scoring framework, scale of inputs or deliverables, the language in the ITT – but several felt it is an imperfect substitute for meaningful dialogue. The very fact of having to guess commissioners' true requirements would make some contractors '*strike out*' some methods as being too risky to their bid. Moreover, normal practice in the department is to not specify a budget for the evaluation, which was widely criticised. Because contractors do not know what the price 'floor' is (i.e. the likely lowest price that other competitors will propose) they cannot take an informed view on how much ground they would have to make up on a superior technical score. This factor, together with guessing what a client means by 'creative', was commonly mentioned as something to deter risk-taking on possibly more costly but also more useful methods. This is especially the case where there is a high weighting for price in the scoring criteria, where a high technical score is unlikely to offset a losing financial score.

Tender assessment – High weighting for price deters creativity and risk taking. Here, interviewees focused mostly on scoring frameworks, and how those are applied, as the main barrier to selecting innovative methods. A significant influence on contractors' decisions about methods at bidding stage was clearly how bids would fare in the scoring exercise compared to competitors. Commissioners sometimes complained about having to compromise with Procurement over price versus technical weightings for scores. It was consistently felt that a high percentage of the total marks awarded to price (more than 30% or 50% were mentioned) will tend to encourage conservative methods choices.

Several observations were made (by contractors and commissioners) that value for money assessments, when using cost per unit metrics, tended to favour quantity of deliverables (e.g. survey respondents) over quality of evaluation outcomes. Innovative methods that require, for example, extensive or ongoing stakeholder involvement, or large amounts of senior resource (e.g. realist approaches), would typically score low on value for money and would be penalised. There were several complaints (including from commissioners) that separating the value for money assessment (often done by procurement) from the technical assessment (often done by the evaluation manager) can prevent a properly contextualised measure of the value of the overall proposal.

Contract management – Lack of flexibility for variation and responding to emergence. Interviewees were often critical of the inflexible pathway for project delivery that tends to be imposed by procurement rules, where contractor compliance is determined in narrow ways, often through quantitative Key Performance Indicators in the contract. Those who spoke about evolutionary or *'recursive'* and iterative approaches, felt these were often at odds with standard government research contracts. Others noted project examples where it had been difficult to change direction in an evaluation in response to learning, because of strict milestone and deliverable requirements.

Procurement has an interest in not allowing too much variation in contract scope, and budget especially, because that could invalidate the original tender competition. But for contractors it increases delivery and resource risk, and it is difficult for commissioners to manage evaluations effectively where uncertainty is intrinsic to the approach. Interviewees consistently called for more flexible approaches to contract management where complexity is involved, to reduce risks of contract failure or deterring contractors from innovating.

Perceptions of intrinsic features of complexity-appropriate approaches and methods that make them difficult to adopt

A range of additional commissioning challenges was identified in relation to intrinsic characteristics of complexity-appropriate methods, including

- How to achieve flexibility in procurement and contracts to accommodate unpredictability and emergence: ' . . . accepting that it's going to be emergent and will need regular reviews of the methodology, of the data, to really understand whether you're getting the right data, using the right methodology'.
- How to enable collaborative working, which is essential to some of the approaches and methods (such as systems mapping).
- How to accommodate multi-stakeholder perspectives and involvement.
- Concerns about cost and timeliness.
- Concern about the usability of findings.

Figure 5 summarises barriers that were identified by interviewees, organised according to the different stages of the commissioning chain. These are often special cases of the general barriers to methods innovation identified in the previous section. A more fundamental challenge, mentioned by a few, would be a need to shift evaluation funders' and users' mindsets from a linear to systems-based way of thinking about policy delivery and accountability.

Supporting the uptake of complexity-appropriate methods – Solutions suggested by interviewees

Changes that could be achieved within the existing procurement framework and process. Interviewees largely felt there is some scope to accommodate complexity-appropriate methods within the present commissioning system. That could include learning from other departments that have used non-standard evaluation methods more extensively (including developmental evaluation approaches). Figure 6 summarises actions that were suggested for different stages in the evaluation commissioning process.

There was some scepticism that these changes could happen easily: commissioners tended to feel they have little influence over the procurement system. Barriers included constraints on individuals' time to influence change outside their 'day job', short time horizons for running procurement exercises, embedded practice and preference, risk aversion, and lack of shared understanding between researcher-commissioners and procurement officials, about the implications of methods choices for evaluation outcomes and their usefulness to policy.[5]

A more radical overhaul of procurement – Move away from 'one size fits all'. A less widely held view was that only a radical overhaul of government research commissioning would enable commissioners to select the best approaches in evaluation studies. Those holding this view wanted a procurement process that was simpler to operate, more flexible in how requirements can be specified and how bids are evaluated (including value for money), and quicker to execute (in terms of the approvals chain). Their priority is to enable commissioning to be more responsive to fast-moving policy.

Pre-specification	Specification and tender assessment	Contract management
Perception that the evolutionary nature of some complexity-informed approaches will not be able to deliver "answers" within tight policy timeframes or offer guarantees on when results will be available. Concern about the usability of findings - whether complexity-appropriate methods can: • deal adequately with questions of attribution • deliver clear-cut conclusions, or whether outcomes will be too "woolly" or nuanced for clients (internal or external) – including in multi-stakeholder evaluations • related to both points, whether the methods offer an equivalent to counterfactual approaches (e.g. where funders need 'simple' answers) Perceptions or experience that some methods are intrinsically more expensive and/or need a longer timeframe than well-known alternatives.	Perception that some complexity-appropriate approaches are intrinsically at odds with a rigid selection process that requires methodologies to be turned into: • a set of tangible outputs and time-bound milestones • to be delivered in a linear progression • and against which contractor performance will be monitored A belief that the flexibility in 'deliverables' that may be required in complexity-appropriate methods would typically be discouraged and/or penalised in tender responses. Concern about how the process evaluates relative 'value for money' of conventional and alternative methodologies, notably in comparing tangible and intangible value (e.g. surveys versus participatory and co-creation processes). A belief that procurement officials may not understand how value is created in complexity-appropriate approaches and therefore (unintentionally) create bias in the tender scoring process in favour of traditional methods.	The scale and nature of flexibility to deal with emergence (scope, budget, timelines, 'deliverables') can't be anticipated at tender stage, which introduces risks commissioners might seek to avoid, namely: • inability to deliver to the contract specification leading to contract failure. • very large contract variations, which would render procurement void, forcing a re-tender exercise. Contractors face equivalent risks of contract-failure and/or resource over-runs (both sides mentioning that resource risk is borne mainly by contractors). Where tender specifications indicate little flexibility in deliverables or costs (e.g. strict KPIs, milestones, deliverables) contractors may avoid specifying complexity-appropriate methods. Perception that procurement rules actively deter collaboration – e.g. in how they describe contract management requirements and monitoring. Extra resource needed for both commissioners and contractors to support collaborative working could be unavailable to research manager commissioners & penalise contractors in price assessments.

Figure 5. Potential barriers arising from intrinsic features of complexity-appropriate methods cited by interviewees.

Pre-specification	Specification and tender assessment	Contract management
Conduct scoping studies where policy is complex - to narrow options and help commissioners devise appropriate specifications and scoring criteria. • Scope delivery and contract risks, as well as methods and costs. • Use internal resource, a panel, or commissioned studies. • Allow time for thinking and iteration – build it early into policy and avoid 'bolted on' evaluation that limits choice of approach and methods. Early and more open dialogue with contractors, to access their knowledge of other methods, narrow down options, and avoid the "apples and pears" issue when 'creativity' is invited in an ITT. Options under existing rules could include: • Staged procurement • Information days • Bidder interviews Alternatively, explore the scope for using collaborative procurement models[1] to enable meaningful dialogue during the tendering process with individual contractors. Existing approaches (e.g. group information days, or tender clarification questions where answers are shared with all contractors) are a weak means of 'consultation' because contractors are reluctant to reveal their ideas to competitors.	Use outcome-focused or less prescriptive ITTs, with enough information for contractors to make informed decisions about 'creative approaches' – including a budget guide (e.g. a range). Where relevant (and proportionate to the budget) invite options and adapt tendering to enable that (e.g. more response space, more time); avoid penalising contractors in the financial scoring for doing so. Rethink the focus of ITTs on deliverables, milestones and tangible outputs, and financial scoring based on cost per unit output. Reward ways of working, insight and outcomes as well as tangible 'deliverables'. Consider other markers of assurance, quality and ability to deliver. Reflect that in equal or higher weight for technical (quality) scores over financial scores. Where the precise scale of specific tasks or outputs cannot be fully anticipated, base scoring for these elements on contractor day rates. Include a post-award scoping stage, where this is of value to the type of approach being proposed, so that contractors proposing it are not penalised in the financial scoring. Review the appropriateness and usability of standardised response templates/platforms for demonstrating fully the potential of a 'new' approach or methods.	Enable greater flexibility to support the use of complexity-appropriate methods. For example: • Using measures that are already available to commissioners, including: a post-award scoping stage to finalise methodology; interim stage gates for review of outcomes and activities; and allowing flexible deployment of resource within identified stages of evaluation projects. • KPIs for contract delivery that are appropriate for emergent methodologies, that enable deliverables and milestones to shift where that is justified by ongoing learning in the project rather than a result of poor performance • A new, more open, agile and collaborative approach to contract management, including: – On-going dialogue rather than intermittent reporting against milestones – Shared learning between commissioner and contractor – Responsive resourcing within overall budget envelopes – Sharing risk, managed through a live risk register to provide transparency and accountability – Active management by the commissioning research manager – more resource than in traditional evaluations, supportive line management to allow flexibility in time-use – Development of new research manager competencies – e.g. being comfortable with uncertainty and adaptability under time and delivery pressure.

Figure 6. Potential solutions for addressing barriers in the evaluation commissioning process to the uptake of complexity-appropriate methods.

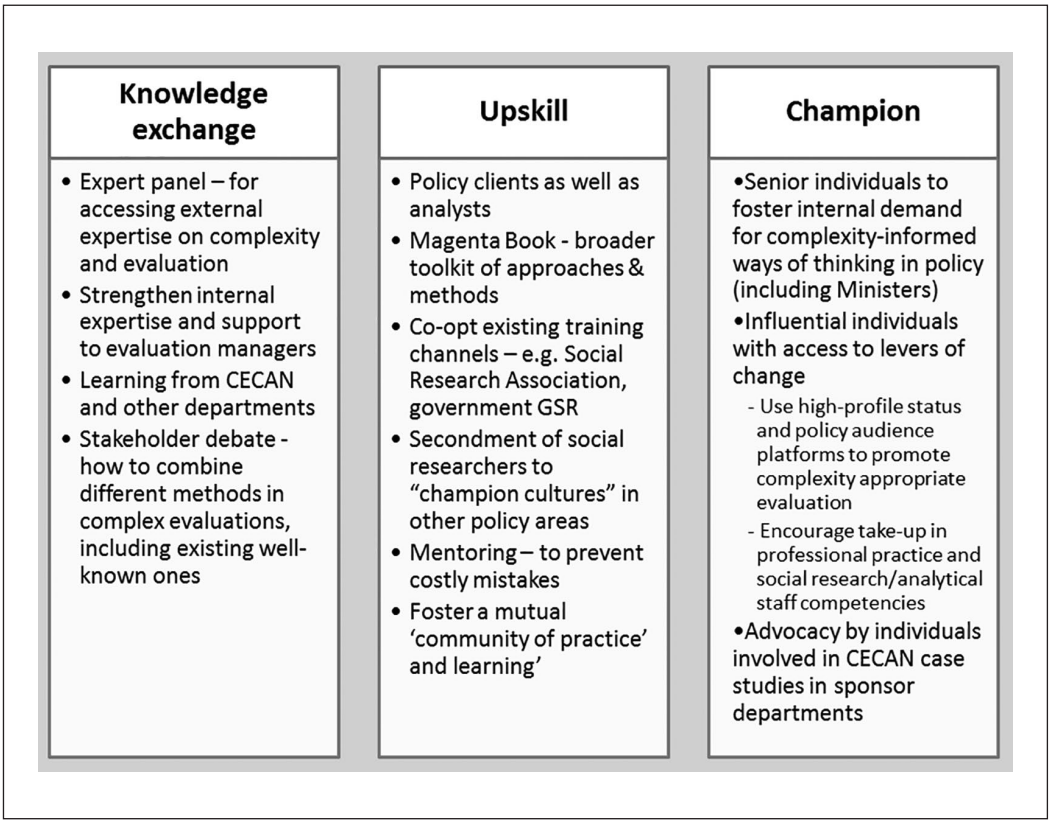


Figure 7. Interviewee suggestions for creating an enabling environment for the uptake of complexity-appropriate evaluation.

Create an enabling environment – Foster demand from higher levels of policy. There was a widespread view that removing procurement barriers needs to be matched by greater demand for complexity-appropriate evaluation at higher levels of policy if its uptake is to increase. That would include Ministers. It would also require a shift from linear to systems-informed ways of thinking about evaluation questions and using evaluation findings. Greater acceptability at higher policy levels would help to de-risk the use of complexity appropriate methods for commissioners and contractors: '*. . . there is something to be done there at a level above where the ITT gets issued in terms of actually saying* "*We are open to these things, these are legitimate methods that should be proposed that we actively encourage*".' Some thought that advocacy by influential internal champions (e.g. chief scientist, social scientists, economists, deputy directors, ministers) is needed, as happened with the adoption of RCTs in policy evaluation. Diffusion of knowledge and skill throughout the department would also be required. Mechanisms for delivering the suggestions in Figure 7 would need to be researched further.

At the level of individual evaluation commissions, various suggestions were made that would support a 'business case' to help de-risk and '*sell*' such approaches to evaluation funders and policy end-users. Interviewees wanted evidence on a range of aspects, summarised in Figure 8. It would need to outline *what* the methods would deliver and at what cost, as much as *how* they work technically: '*. . . [funders] want to know the methods we use are reliable*

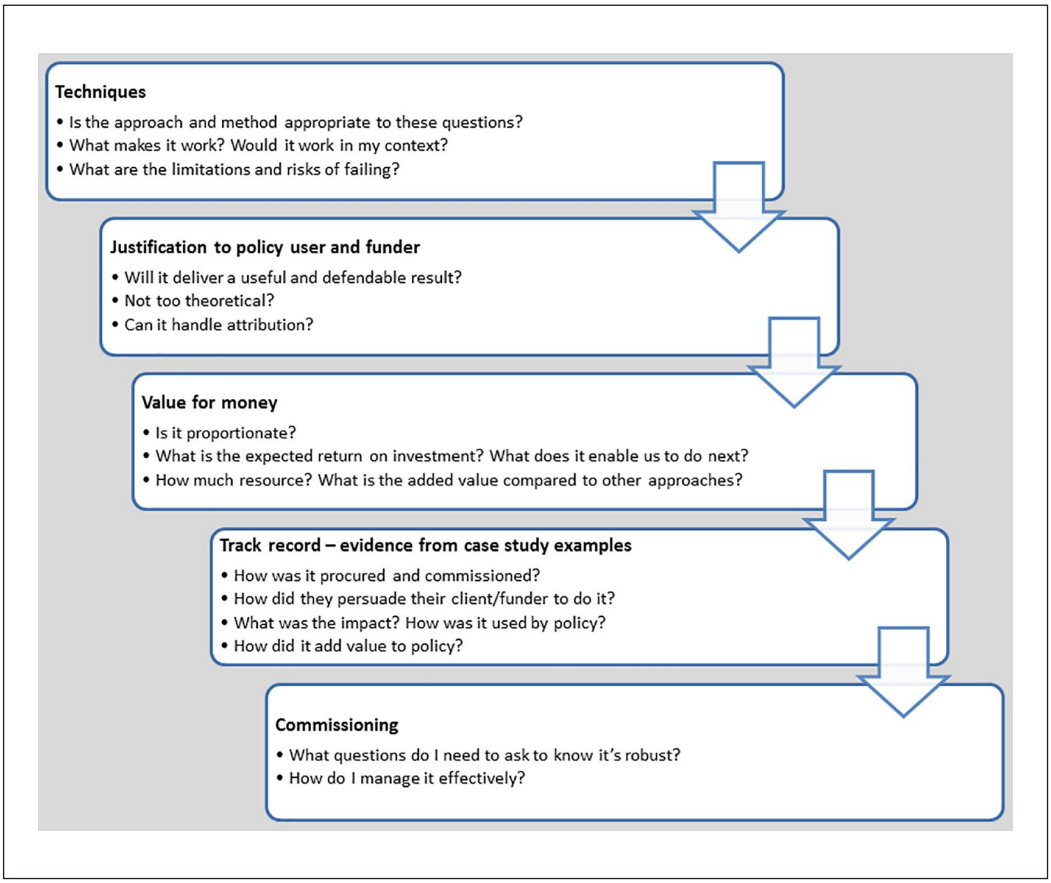


Figure 8. Interviewee suggestions on the evidence needed to support a case to evaluation funders and policy managers to commission complexity-appropriate methods.

and credible, and robust, but beyond that most of our managers will glaze over when you get into the technicalities of it . . .'. Some suggested that having better evidence on cost and performance would make it easier to score the relative risk and value of such methods in tender assessments.

Interviewees gave examples where they felt there were opportunities to build the evidence: by trialling methods on small and low-profile projects, where the consequences of failure would be low; or in new policy areas, where there is less evaluation history and embedded methodological preference.

Conclusion

This paper addressed gaps in the literature on barriers to the commissioning of complexity-appropriate evaluation methods and how the commissioning system could be improved. Unusual in the literature (Schneider et al., 2016), the research examined perspectives of both evaluators (research contractors) and commissioners operating in the same policy

space (a UK government department), drawn from interviews with 9 commissioners and 10 contractors.

We theorised that approaches and methods selected in commissioned evaluations evolve from the behaviour of actors on both 'sides' of competitive tendering, interacting with a rules-based procurement process. Behaviours are further influenced and shaped by the contexts in which the actors are operating, including the political and intellectual zeitgeist. The tender competition is the decision-making nexus where rules and behaviours interact and where the wider influences on commissioners and contractors are crystallised in the choices that are eventually made about approaches and methods.

The study found that behaviours of commissioners and contractors alike were shaped not only by the practical constraints of procurement processes but also the wider policy and organisational context. It confirmed that higher-level commissioning barriers emerge from the political contexts where evaluation demand arises (e.g. Rutter, 2013; Schneider et al., 2016; The LSE GV314 Group, 2014). Those higher-level barriers included tight budgets and timeframes, evaluation being embedded to different levels in different policy areas, lack of knowledge and skills about complexity-appropriate evaluation, concern over anticipated results and political influence. There was consensus across contractors and commissioners that the shape of demand from policy clients is a crucial influence on choices about evaluation approach and methods. Scientific relevance and quality objectives are important but so is making sure that results *'land'* with policy clients – that is, they are useful for the purpose the client has in mind, which may be a political one (as in The LSE GV314 Group, 2014).

Like Rutter (2013), who found that avoiding political risk was a barrier to the wider uptake of evaluation in evidence-based policy making, we found it was similarly a constraint on the choice of methods when evaluation tenders are being specified. Indeed, risk in many forms emerged as a crucial influence on decisions made about methods during the whole commissioning chain, from securing approval and budget for an evaluation all the way through to communicating the findings to policy users. For commissioners and contractors, it spanned perceptions of technical risks (whether methods would work), delivery risks, contract risks, organisational risks and personal risks (e.g. to their own performance and professional standing). Risk for contractors further revolved around having to guess commissioners' true requirements, including budgets, and of being non-compliant or uncompetitive if they got it wrong.

On both sides, controlling for risk often meant conservative choices about evaluation methods, in favour of 'standard' approaches that proposers could be confident are acceptable to clients. Prevailing evidence cultures and preferences across government were an important influence here. In addition, combined pressures to raise the quality of evaluation practice and to *'do more with less'* budget were said to favour risk averse behaviours and sticking with well-known and accepted methods (e.g. counterfactual impact methods). Equally, some pointed to an opportunity for developing a more plural *'evaluation toolbox'* where commissioners or policy are finding that *'standard'* methods are not suited to the complexity of the questions they are trying to answer.

Turning to the rules and mechanisms of the tendering process itself, the study identified changes that might be made at key points to make it easier for novel methods to compete effectively against traditional evaluation methods. Interviewee observations about barriers and solutions were described in Figures 5 and 6. From those, we have drawn out in Figure 9 some principles for adapting existing procurement models to be more 'complexity-friendly'.

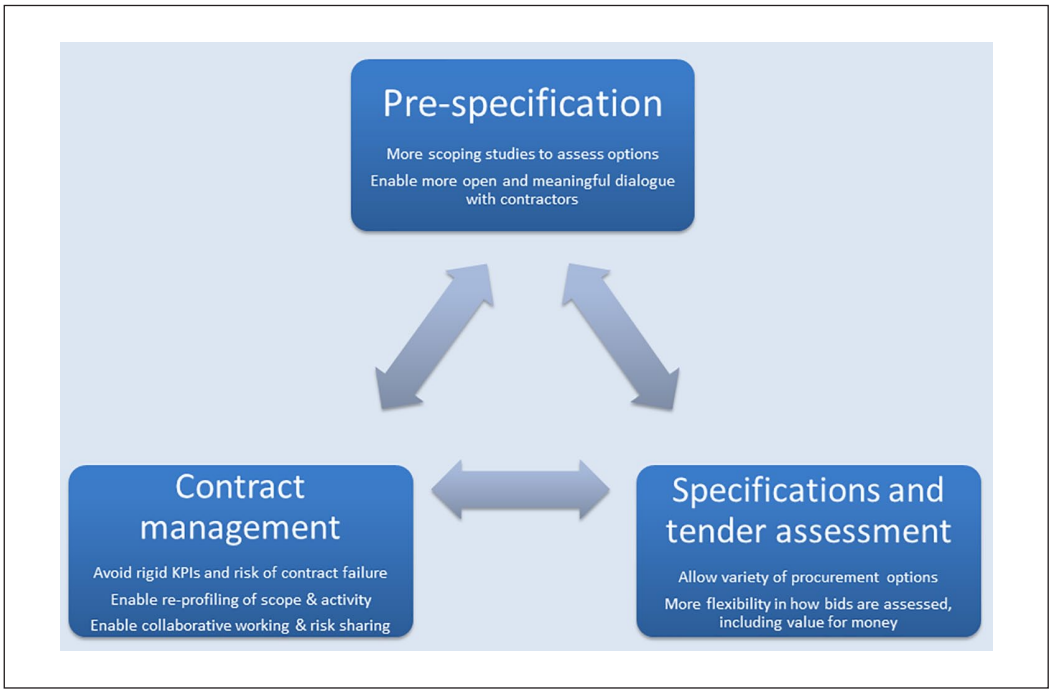


Figure 9. Principles for reforming key aspects of evaluation procurement to enable commissioning of complexity-appropriate methods, drawn from interviewees' detailed suggestions.

Interviewees who were familiar with complexity-appropriate methods identified some intrinsic features that are especially difficult to accommodate in existing evaluation procurement processes. These relate to

- Flexibility of scope, tasks and resource allocation – to accommodate emergence in the parameters being evaluated and responsive evaluation frameworks and tasks
- Co-productive ways of working – commissioner-contractor relationships based on trust rather than command-and-control, mutual learning, some shared risk taking, and resource to support the active management and ongoing dialogue needed to make it effective

While many interviewees felt that existing procurement mechanisms could be revised to make room to commission complexity-appropriate methods, some were sceptical and thought that a more radical overhaul of research and evaluation procurement is needed. That might include modifying the conventional relationship between commissioner and contractor (Knight et al., 2017) to allow greater interaction, collaboration, shared goals and risk sharing.

Changing mindsets and fostering demand for complexity-appropriate evaluation at higher-levels of policy was also called for. Without that, revising the evaluation procurement process would fail to encourage the wider uptake of complexity-appropriate methods, according to some. Suggestions were made for upskilling and knowledge exchange in government research and policy professions, credible and high-profile champions to build support internally, and for commissioner-contractor joint communities of practice. A need to generate evidence to

answer commissioner and policy questions about cost, relevance and reliability of complexity-appropriate methods was also flagged.

We could therefore conclude that reforming procurement needs to be matched by a shift in demand at higher levels in policy (including Ministers), to give those lower in the hierarchy the confidence and 'permission' to procure complexity-appropriate evaluation and, for contractors, to offer it. Making the changes to tendering processes that could support uptake would require a lead from procurement functions in government: for example, to explore alternative commissioning models that enable more collaborative approaches than are possible in conventional competitive tendering and contract management. Models like this may exist[6] but were outside the scope of this research. On the demand-side, further research into policy makers' attitudes towards, and appetite for, complexity-appropriate evaluation is needed.

Acknowledgements

The authors gratefully acknowledge the contributions of all the interviewees who generously gave their time. The views reflected in the findings are the authors' interpretation and do not represent the view of any individual or organisations they are affiliated with.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Economic and Social Research Council (grant numbers ES/N012550/1 and ES/S000402/1).

ORCID iD

Pete Barbrook-Johnson https://orcid.org/0000-0002-7757-9132

Notes

1. Research procurement was subject to EU rules at the time of the research (2018).
2. Interviewees often used the terms 'approach' and 'methods' interchangeably, so no clear distinction is made in this paper. The term 'complexity appropriate methods' is used to encompass both evaluation approaches *and* evidence gathering methods that are informed by complexity science.
3. 'New' is used throughout to describe methods that are new in an innovation sense – either new to the world or new in this application or context.
4. As noted in the literature review, since this research was completed UK government published a Magenta Book annex on complexity, to support uptake.
5. But noting this is a one-sided view because procurement officials were not interviewed for this research.
6. Collaborative procurement was not investigated within this research, but it was suggested in some interviews and at project workshops as an area worth exploring. Provisions in updated EU procurement regulations in 2014 may offer scope for more flexible consultation with the market and procedures for situations where services are entirely new. Expert advice would be needed to explore whether these could be used in evaluation procurement. See: Procurement Policy Note: Availability of Procurement Procedures (Decision Tree) https://www.gov.uk/government/publications/procurement-policy-note-1215-availability-ofprocurement-procedures-decision-tree.

References

Anderson S, Allen P, Peckham S, et al. (2008) Asking the right questions: Scoping studies in the commissioning of research on the organisation and delivery of health services. *Health Research Policy and Systems* 6(7): 1–2.

Barbrook-Johnson P, Proctor A, Giorgi S, et al. (2020) How do policy evaluators understand complexity? *Evaluation* 26(3): 315–32.

Barbrook-Johnson P, Schimpf C and Castellani B (2019) Reflections on the use of complexity-appropriate computational modeling for public policy evaluation in the UK. *Journal on Policy and Complex Systems* 5(1): 55–70.

Barnes M, Matka E and Sullivan H (2003) Evidence, understanding and complexity. *Evaluation* 9(3): 265–84.

Boehnert J, Penn A, Barbrook-Johnson P, et al. (2018) The visual representation of complexity. *CECAN*. Available at: https://www.cecan.ac.uk/resources

Broer T, Bal R and Pickersgill M (2017) Problematisations of complexity: On the notion and production of diverse complexities in healthcare interventions and evaluations. *Science as Culture* 26(2): 135–60.

Davoudi S, Harper G, Petts J, et al. (2015) Judging research quality to support evidence-informed environmental policy. *Environmental Evidence* 4(1): 9.

De Laat B (2013) Evaluator, evaluand, evaluation commissioner: A tricky triangle. In: Loud ML and Mayne J (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. Thousand Oaks, CA: SAGE, 15–36.

De Laat B and Williams K (2013) Evaluation Use within the European Commission (EC): Lessons for the evaluation commissioner. In: Loud ML and Mayne J (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*. Thousand Oaks, CA: SAGE, 147–74.

Defra (2020) Complexity evaluation framework. Available at: http://sciencesearch.defra.gov.uk/Default.aspx?Menu=Menu&Module=More&Location=None&Completed=220&ProjectID=20401

Gates EF (2016) Making sense of the emerging conversation in evaluation about systems thinking and complexity science. *Evaluation and Program Planning* 59: 62–73.

Gates EF (2017) Learning from seasoned evaluators: Implications of systems approaches for evaluation practice. *Evaluation* 23(2): 152–71.

Giorgi S (2017) How to improve the evaluation of complex systems to better inform policymaking learning from evaluating Defra's reward & recognition fund. *CECAN Report*. Available at: https://www.cecan.ac.uk/resources

Government Social Research Profession (2018) The government social research code – People and products. Available at: https://www.gov.uk/government/publications/the-goverment-social-research-code-people-and-products

Head BW (2010) Reconsidering evidence-based policy: Key issues and challenges. *Policy and Society* 29(2): 77–94.

HM Treasury (2020) Magenta Book Annex: Handling complexity in policy evaluation. Available at: https://www.gov.uk/government/publications/the-magenta-book

Knight AD, Lowe T, Brossard M, et al. (2017) A whole new world – Funding and commissioning in complexity. *Collaborate CIC*. Available at: https://collaboratecic.com/a-whole-new-world-funding-and-commissioning-in-complexity-12b6bdc2abd8

McIntosh EJ, Chapman S, Kearney SG, et al. (2018) Absence of evidence for the conservation outcomes of systematic conservation planning around the globe: A systematic map. *Environmental Evidence* 7(1): 1–22.

Mowles C (2014) Complex, but not quite complex enough: The turn to the complexity sciences in evaluation scholarship. *Evaluation* 20(2): 160–75.

National Audit Office (2013) Evaluation in government. Available at: https://www.nao.org.uk/report/evaluation-government/

Puttick R (2018) Mapping the standards of evidence used in UK social policy. *Alliance for Useful Evidence*. Available at: https://www.alliance4usefulevidence.org/publication/mapping-the-standards-of-evidence-used-in-uk-social-policy/

Reynolds M, Gates E, Hummelbrunner R, et al. (2016) Towards systemic evaluation. *Systems Research and Behavioral Science* 33(5): 662–73.

Rutter J (2013) Evidence and evaluation in policy making: A problem of supply or demand? *Institute for Government*. Available at: https://www.instituteforgovernment.org.uk/publications/evidence-and-evaluation-policy-making

Sanderson I (2000) Evaluation in complex policy systems. *Evaluation* 6(4): 433–54.

Schneider CH, Milat AJ and Moore G (2016) Barriers and facilitators to evaluation of health policies and programs: Policymaker and researcher perspectives. *Evaluation and Program Planning* 58: 208–15.

Schwarzman J, Bauman A, Gabbe B, et al. (2018) Organizational determinants of evaluation practice in Australian prevention agencies. *Health Education Research* 33(3): 243–55.

Schwarzman J, Bauman A, Gabbe BJ, et al. (2019) Understanding the factors that influence health promotion evaluation: The development and validation of the evaluation practice analysis survey. *Evaluation and Program Planning* 74: 76–83.

Stame N (2010) What doesn't work? Three failures, many answers. *Evaluation* 16(4): 371–87.

Stern E (2015) Impact evaluation: A guide for commissioners and managers. *Bond*. Available at: https://www.bond.org.uk/resources/impact-evaulation

The LSE GV314 Group (2014) Evaluation under contract: Government pressure and the production of policy research. *Public Administration* 92(1): 224–39.

Walton M (2014) Applying complexity theory: A review to inform evaluation design. *Evaluation and Program Planning* 45: 119–26.

Walton M (2016) Expert views on applying complexity theory in evaluation: Opportunities and barriers. *Evaluation* 22(4): 410–23.

Williams B (2015) Prosaic or profound? The adoption of systems ideas by impact evaluation. *IDS Bulletin* 46(1): 7–16.

Jayne Cox is a CECAN Fellow and Director of Brook Lyndhurst, a sustainability-behaviours research company. She has worked on both sides of the commissioning divide for UK public bodies.

Pete Barbrook-Johnson is a Senior Research Fellow in the Department of Sociology at the University of Surrey, a UKRI Innovation Fellow, and a member of CECAN.