Journal of MultiDisciplinary Evaluation
Volume 16, Issue 37, 2020

ISSN 1556-8180
http://www.jmde.com

Learning Lessons for Evaluating Complexity Across the Nexus: A Meta-Evaluation of Environmental Projects

William R. Sheate
Collingwood Environmental Planning (CEP), London
Imperial College London Centre for Environmental Policy, London

Clare Twigger-Ross
Collingwood Environmental Planning (CEP), London

Liza Papadopoulou
ICF Consulting, London (formerly at CEP)

Rolands Sadauskis
Collingwood Environmental Planning (CEP), London

Owen White
Collingwood Environmental Planning (CEP), London

Paula Orr
Collingwood Environmental Planning (CEP), London

Ric Eales
Collingwood Environmental Planning (CEP), London

Background: A major gap in environmental policy making is learning lessons from past interventions and in integrating the lessons from evaluations that have been undertaken. Institutional memory of such evaluations often resides externally to government, in evaluation practitioner contractors who undertake commissioned evaluations on behalf of government departments.

Purpose: The aims were to learn the lessons from past policy evaluations, understand the barriers and enablers to successful evaluations, to explore the value of different types of approaches and methods used for evaluating complexity, and how evaluations were used in practice.

Setting: A meta-evaluation of 23 environmental evaluations undertaken by Collingwood Environmental Planning Ltd (CEP), London, UK was undertaken by CEP staff under the auspices of CECAN (the Centre for Evaluation of Complexity Across the Nexus – a UK Research Councils funded centre, coordinated by the University of Surrey, UK). The research covered water, environment and climate change nexus issues, including evaluations of flood risk, biodiversity, landscape, land use, climate change, catchment management, community resilience, bioenergy, and European Union (EU) Directives.

Intervention: Not applicable.

Research design: A multiple embedded case study design was adopted, selecting 23 CEP evaluation cases from across a 10-year period (2006-2016). Four overarching research questions were posed by the meta-evaluation and formed the basis for more specific evaluation questions, answered on the basis of documented project final reports and supplemented by interviews with CEP project managers. Thematic analysis was used to draw out common themes from across the case categories.

Findings: Policy context invariably framed the complex evaluations; as environmental policy has been spread beyond the responsibility of government to encompass multiple stakeholders, so policy around nexus issues was often found to be in a state of constant flux. Furthermore, an explicit theory of change was only often first elaborated as part of the evaluation process, long after the policy intervention had already been initiated. A better understanding of the policy context, its state of flux or stability as well as clarity of policy intervention's objectives (and theory of change) could help significantly in designing policy evaluations that can deliver real value for policy makers. Evaluations have other valuable uses aside from immediate instrumental use in revising policy and can be tailored to maximise those values where such potential impact is recognised. We suggest a series of questions that practitioners and commissioners could usefully ask themselves when starting out on a new complex policy evaluation.

Keywords: *evaluation; complexity; policy use; natural environment*

Introduction

Evaluation is now seen as an integral part of policy making (HM Treasury, 2018, 2020)—evaluating how interventions are being or have been implemented, whether they have been effective in delivering what was intended, what unforeseen impacts there might have been and how best to revise or refine policy in light of those findings. Increasingly environmental policy in particular is seen to be inherently complex, uncertain and long-term (European Environment Agency, 2011) because of the multi-faceted nature of the environment as well as the multitude of stakeholders involved, and uncertainty around the impacts of human activity and policy interventions over the long time periods needed to understand and address such impacts.

A major gap in policy making has been learning the lessons from past interventions and integrating the lessons from evaluations that have been undertaken into future interventions. All too often there is little transparency about what happens to evaluations—how they are used, if at all, by policy makers. Some policy interventions have a short life-span (e.g., 2-3 years), are evaluated, but then dropped, perhaps because of limited funding availability, perhaps even because of the outcome of the evaluation, although that may never be made public. Follow-up in terms of evaluations and their impact is also difficult because of the turn-over of civil servants within government departments and the associated loss of institutional memory. Such memory often resides externally to government, in evaluation practitioner contractors who undertake commissioned evaluations on behalf of government departments, contractors like Collingwood Environmental Planning (CEP), based in the UK in London, for which the authors of this paper work.

This issue of evaluation use or impact was of particular interest because the evaluation studies undertaken by CEP could all be characterized as evaluations of complex nexus [1] policy related interventions or initiatives. This paper investigates the range of uses and shows how evaluations can usefully impact in different parts of the policy system beyond the traditional instrumental use. The paper draws on a meta-evaluation—an evaluation of evaluation studies carried out by CEP, under the auspices of CECAN [2] (the Centre for Evaluation of Complexity Across the Nexus—a UK Research Councils funded centre, coordinated by the University of Surrey, UK). A sample of 23 projects was selected on the basis of publicly available final evaluation reports and current CEP staff with knowledge of those projects. The research covered water, environment and climate change nexus issues, including evaluations of flood risk, biodiversity, landscape, land use, climate change, catchment management, community resilience, bioenergy, and EU Directives.

Background

Defining Complexity in Relation to Policy Interventions/Initiatives at the Nexus

The issue of complexity has recently been addressed in a new update and supplementary guide to the UK Government's The Magenta Book (HM Treasury, 2020 a, b).

Complexity can be defined as:

> Key characteristics of complex systems include: adaptation to changes, feedback loops, multiple scales, thresholds for change, areas of relatively high and low stability, past states influencing possible future states, being highly dynamic and being an open system, impossible to bound. These result in complex systems both social and ecological, exhibiting tipping points, emergent new properties and unpredictability (CECAN, 2018).

Three different aspects of complexity can be identified (HM Treasury, 2020b; HM Treasury, 2018; Jaffe et al., 2005):

[1] The 'nexus' in this context is taken to mean the focus on interconnectedness across environmental domains specifically those of flooding, land use, climate change, catchment management, and biodiversity.

[2] https://www.cecan.ac.uk/

- The complexity of the problem/issue that the intervention being evaluated is trying to address;
- The complexity of policy response being evaluated;
- The complexity of impacts of the intervention being evaluated.

These three types of complexity can be expanded upon (HM Treasury, 2020 b); Table 1 presents examples of the different aspects of complexity.

Table1
Examples of Types of Complexity

Problem-related complexity	▪ Problem has multiple elements ▪ Variability in the physical / environmental characteristics of the area / location ▪ Geographic spread / scale of the problem ▪ Sensitivity to socio-demographic characteristics of the area / target population Level of unpredictability in the problem
Policy/Response-related complexity	▪ Multiple components / elements included in the policy/programme/initiative ▪ Multiple agencies / actors / stakeholders involved or targeted by the policy (may include conflicting interests) ▪ Degree of flexibility or tailoring / changes in the policy during implementation ▪ Geographic spread/ scale of the policy response ▪ Competing / interacting policies (at a UK or EU level)
Impact-related complexity	▪ Multiple types / range of possible / expected outcomes and impacts ▪ Unexpected / unintended impacts (positive or negative) ▪ Interactions between components of a policy ▪ Timescales over which impacts might occur

Some of these might be better classified as making the situation "complicated" rather than "complex" per se. The difference between these two draws on the distinction between what is 'complicated' (multiple components) and what is 'complex' (uncertain and emergent) (Glouberman & Zimmerman, 2002). As Rogers (2008) notes, these concepts have been adopted by a number of authors (Downe et al., 2012; Rogers, 2008; Snyder, 2013). The distinction, explained in an evaluation context, can be clarified as (Rogers, 2008):

- Complicated project/policy and/or evaluation theory: Elements that are inherent to the project or policy design, including multiple components, multiple actors/stakeholders, multiple and diverse activities, multiple simultaneous and/or alternative causal strands.
- Complex evaluation/programme theory: Complexity refers to recursive causality (with reinforcing loops), disproportionate relationships (where at critical levels, a small change can make a big difference—a 'tipping point') and emergent outcomes.

Both have an impact on evaluation and this paper draws out ways in which that is the case.

Defining Policy Evaluation and its Objectives

The Magenta Book (HM Treasury, 2020a, p. 9) describes policy evaluation as "the systematic assessment of a Government policy's design, implementation and outcomes. It involves understanding how a government intervention is being or has been implemented and what effects it has had, for whom and why. It also comprises identifying what can be improved and how, as well as; estimating overall impacts and cost-effectiveness." In practice, these questions and their responses are much more complex including considerations of how different features of the policy affected the way it performed and delivered, and how its outcomes varied across those it impacted upon: what worked for whom in what circumstances (HM Treasury, 2020a).

The overarching objective of evaluation is to offer an unbiased assessment of a policy's performance by measuring outcomes and impacts in order to assess whether the anticipated benefits of a policy have been realised. A good evaluation, however, does not stop there, but ensures that lessons are learned and communicated so that they may inform future proposals and policies. It therefore provides information on what could be improved in the design and delivery of a policy. In doing so it often involves an evaluation of the process of policy implementation as one of the factors influencing success.

The purpose of the evaluation depends on 'what' is being evaluated and 'when' or in which stage of the policy cycle or the policy design and implementation process it is being carried out. An evaluation that takes place alongside the policy's implementation can support the delivery of the policy by identifying what works well or less well and why. In doing so, it can help improve the effectiveness of the policy in meeting its objectives, while it also offers the opportunity for course correction if necessary. An evaluation taking place following the policy delivery can allow lessons to emerge that will inform the development of new policies.

This assumes that policy making follows a traditional (ROAMEF)[3] 'policy cycle' and that evaluation occurs as a specific stage in that cycle (e.g., as understood in The Magenta Book). It is also widely recognised that such a policy cycle is in practice often very fuzzy. Evaluation needs to recognise the fuzziness of the policy process because it affects if, how and when evaluation might have any influence on policy. Hallsworth (2011) notes, as a result of decentralisation and devolution of responsibilities in policy making from the centre, that increasingly:

- Policy formulation and implementation are not separate, but intrinsically linked;
- The potential outcomes of the policy itself may change significantly during implementation;
- Complexity in public service systems often means central government cannot directly control how these changes happen;
- The real-world effects policies produce are often complex and unpredictable.

This means that in the UK policy making no longer (if it ever did) necessarily follows a typical policy cycle where evaluation is part of the cycle that leads to refinement of the policy. Instead, policy making and implementation are now seen as part of a more dynamic system, termed 'system stewardship' by Hallsworth (2011). This is a more flexible and adaptive model, but also means that evaluation now faces a 'triangle' of purpose (or policy goal), design (or policy direction) and implementation (or action, realization) simultaneously (rather than being part of a cycle) and needs to be responsive to changing circumstances. Both policy making and its consequences are more complex and so evaluation needs to be responsive to this complexity, as well as complexity intrinsic to the subject matter of environmental policy and the nexus. Conventional performance indicators are often poorly suited to this increasing complexity and uncertainty; more flexible participatory approaches are therefore needed in evaluation that can deal with

[3] Rationale, Objectives, Appraisal, Monitoring, Evaluation, Feedback (HM Treasury, 2018).

multiple perspectives of different stakeholders (Hallsworth, 2011).

Evaluation Use

The term evaluation use or utilisation refers to the way(s) in which evaluations and their findings affect operations, decisions and outcomes (Henry & Mark, 2016; Balthasar, 2009; Kirkhart, 2000) and it is fundamental to demonstrating an evaluation's success. Although the aims and objectives of an evaluation are a good indication of the evaluation use, there are a number of factors influencing the actual impact of an evaluation. According to (Balthasar, 2009), these can include:

- Institutional factors, such as the organisation triggering the evaluation;
- Environmental factors, such as the evaluation culture;
- Process-related factors, such as mechanisms in place for stakeholder engagement or policy implementation.

A key element encompassed in both institutional and environmental factors above, is the human element, referring mainly (but not solely) to the intended users of the evaluation. BetterEvaluation (no date) notes that "the use of an evaluation often depends on how well the report meets the needs and learning gaps of the primary intended users". However, there is also an element of how evaluation is perceived by users. Peck and Gorzalski (2009), like others before them, note that evaluations are often seen by commissioning bodies and organisations as 'ideas for change' rather than concrete improvements to be implemented. Adopting this attitude towards evaluation has a significant impact on how the various recommendations are perceived and to what extent (or whether) these are taken on board.

Looking at theory and practice, Peck and Gorzalski (2009) combine types of change (Downs, 1967, and Johnston, 1988, as cited by Peck & Gorzalski, 2009) and types of influence/use (Kirkhart, 2000) in an integrated framework. Bringing together the different perspectives emerging from theory and practice and adjusting them to fit the context of environmental policies, programmes and initiatives, the meta-evaluation study reported here adopted the following categorisation of evaluation uses (see Table 2):

Table 2
Evaluation Use Categories

Instrumental/Purpose-based use	▪ direct use of an evaluation's findings in decision making or problem solving ▪ suggests changes to overall mission and aims
Conceptual use	▪ suggests changes in thinking or behaviours
Process-based/Structural use	▪ suggests changes on the basis of knowledge gained while undertaking the evaluation ▪ suggested changes may refer to the organisation's or programme's structure
Strategic/Persuasive use	▪ evaluation is used to influence policy ▪ can provide arguments in support of a political position (or not)

In line with earlier comments around attitudes towards evaluation, Peck and Gozalski found that very little instrumental use existed with most evaluation use being conceptual, as only a few organisations studied had implemented specific evaluation recommendations (Peck and Gorzalski, 2009).

Literature on evaluating complexity often refers to programme theory, otherwise known as 'theory of change', 'intervention logic' or use of 'logic models' as an approach that allows evaluators to develop a causal chain between the programme inputs, activities, outputs and intended and observed outcomes (Rogers, 2008; Pawson et al., 2005; Sanderson, 2000). A realist review or realist evaluation, offers a useful model of research synthesis that is designed to work with complex social interventions or programmes (Magro & Wilson, 2013; Pawson et al., 2005; Pawson, 2013; Wong et al., 2014). Grounded in the theory that underpins a programme or intervention it seeks to collect evidence from a diverse range of available sources. The review combines theoretical understanding, empirical evidence, case studies and formal reports with qualitative data from interviews often undertaken with those involved in the evaluations, to explain the relationship between the context in which the intervention is applied, the mechanisms by which it works and the outcomes which are produced. It provides an explanatory analysis aimed at what works for whom, in what circumstances, in what respects and how.

Hargreaves and Podems (2012) also argue that theory-based approaches are more appropriate in dealing with complex interventions as they provide early feedback about what is working or not, and why, thus allowing early intervention and course correction. In situations where there is uncertainty regarding the approach of the programme/project and the expected outcomes, such evaluations (also characterised as developmental)—can prove more appropriate compared to formative and summative evaluations (BetterEvaluation, no date) (see Table 2). Patton (2002) suggests that the increased involvement of key stakeholders in the project delivery, decision-making and monitoring and evaluation is a more pragmatic approach in complex evaluations. Benefits of using such an approach include the development of a sense of ownership for those stakeholders involved who are also able to contribute local knowledge and insights for the evaluation. However, Hargreaves and Podems (2012) warn of challenges in stakeholders expressing contrasting views while others warn of these evaluations being regarded less objective (BetterEvaluation, no date).

Research Approach and Methodology

Introduction

The aims of the research were:

1. To learn the lessons from past policy evaluations;
2. To understand the barriers and enablers to successful evaluations, where success is measured by: (a) Whether the evaluation meets its own objectives; (b) The impact of that evaluation;
3. To explore the value of different types of approaches and methods used for evaluating complexity.

Four overarching research questions were posed by the meta-evaluation and are reported in this paper:

1. Were the evaluations fit for purpose, and was their purpose clear? What lessons can be learnt about assessing the effectiveness of complex policy interventions/initiatives across the nexus?
2. How has the framing of the evaluation been more or less useful for understanding complexity (e.g., logic model, objectives led)?
3. What methods have been used for dealing with aspects of complexity found within environmental policy? Which methods appear to have been most effective? Were some methods and techniques more suited to certain types of complexity?
4. What factors lead to an evaluation being more (or less) influential in policy changes / outcomes / evaluation use?

These were used to structure more specific evaluation questions, answered on the basis of documented project final reports and supplemented by interviews with CEP project managers. A multiple embedded case study design was adopted, selecting 23 CEP evaluation cases from across a 10-year period

(2006-2016), which were categorised as shown in Table 3.

These categories were based on common policy contexts in which the evaluations were taking place, rather than, for example, the type of evaluation (e.g., formative, summative) or the specific nexus sector since many of the cases covered multiple types of evaluation and sector.

Table 3
Multiple, Embedded Case Study Sesign (numbers indicate numbers of cases in each category)

CEP evaluations 2006-2016 (23)	**Policy interventions (6)**	EU policy interventions (EUP) (3) National policy interventions (NP) (3)
	Programme level interventions/initiatives (17)	Programme level policy interventions (PPI) (9) Programme level initiatives (i.e. not linked directly to implementing specific policy) (PI) (8)

Each case was then categorised in a master Excel spreadsheet according to the criteria shown in Table 4. The categories chosen were those which related to the meta-evaluation questions, enabled comparison between the projects and came out of the literature review around complexity and evaluation. The projects chosen for the meta-evaluation were diverse and therefore categories were needed that could describe all the projects.

Table 4
Evaluation Categories Used to Describe Selected Projects for the Meta-Evaluation

Category	Explanation
Scale	Geographical scale: local, regional, national and multi-national
Policy area	The focal policy area of the evaluation
Type of evaluation	The evaluation approach used: *formative, summative, developmental, participatory, theory-based, ex-ante, ex-post, experimental, quasi-experimental, non-experimental*. These are not mutually exclusive categories as some evaluations were combinations of different types.
Data collection methods	Methods used to collect data: *literature review, data/indicator review, observation, surveys/questionnaires, developing case studies, interviews, workshops/events, steering group/expert advice, participant diaries*
Types of complexity	Three areas of complexity were defined: • *issue-related complexity* • *policy/response-related complexity* • *impact-related complexity*
Evaluation use	Four types of use were examined: *instrumental, conceptual, strategic and process-related uses*

Category	Explanation
Budget	*Six bands of budget were included in this category: £20,000; £21,000- £50,000, £51,000-£99,999, £100,000- £199,999, £200,000-£300,000*

Method of Classification

Once the categories had been agreed each of the projects was classified according to those categories by one member of the research team on a presence or absence basis for each category. For ease of use an Excel spreadsheet was developed listing all the projects and the categories. Once the first classification had been carried out each of the CEP project managers was asked to verify the classifications for their projects and to choose which three aspects of complexity were most reflected within their projects, and the types of evaluation use (where known). The classification process was iterative, with new categories being added through discussion with the project board and wider CECAN community. Specifically, the evaluation use category was divided into four sub-categories to capture the different types of use and the budget category was added to give an indication of the range of budgets associated with each evaluation.

Answering the Specific Evaluation Questions

Once the classification process was finished, each of the projects was examined in relation to the specific evaluation questions and their sub-questions (see Table 4) with responses to those questions provided from project documentation, typically published final reports, supplemented by semi-structured interviews internally with the CEP project managers. All responses were recorded in the Excel spreadsheet with a clear distinction made between sources (by using different coloured text). From these individual responses, themes, messages, observations and examples were drawn out for each of the questions across all projects within a case category, for example, for EU policy interventions. These were recorded within the Excel spreadsheet. More work was carried out by two members of the team to draw out key themes which were then used to compare between the case categories.

Table 4
Specific Evaluation Questions and Sub-Questions

Questions and Sub-Questions
1.0 Were the evaluations fit for purpose
1.1 What were the objectives of the policy or interventions? i. Were they clear and appropriate? ii. Was there consensus on the objectives of the intervention?
1.2 What were the objectives of the evaluation? i. Were they clear and appropriate? ii. Was there consensus on the evaluation objectives?
1.3 What were the circumstances within which the evaluation took place? i. Intervention governance arrangements (e.g., national/local tensions etc.)? ii. Stable or evolving policy context (e.g., changes in higher-level political priorities etc.)?

Questions and Sub-Questions
1.4 Project management context (steering group)?
i. Large or small steering group.
ii. Project manager (quality of).
2.0 How was the evaluation 'framed' for complexity (e.g., logic model, theory of change)?
i. Was the evaluation framework developed as part of the evaluation, or provided by the commissioning authority?
ii. Were stakeholders involved in agreeing the evaluation framework?
3.0 What methods were used by the evaluation for dealing with complexity?
i. What method/s were predominantly used?
ii. Why were some methods used / not used? Where the methods appropriate?
4.0 What happened to the evaluation; how was it used?
i. How (if at all) did it inform policy?
ii. Description and explanation, e.g., look across previous questions (e.g., if evolving policy, to what extent if at all did an evaluation influence that change in policy?)

Our approach to thematic analysis was firmly rooted in qualitative research, drawing on grounded theory where key themes are identified and coded from the data. Emerging themes from the spreadsheet of cases were clustered and reviewed by two team members. Common themes across the case study categories were identified, some of which were pre-determined codes derived from the evaluation questions in Table 4 (e.g., clarity of objectives, stability of policy, existence of explicit theory of change); others emerged from the data, for example, strength or weakness of the policy cycle. Given the purpose of the study was to answer the four key evaluation questions from across all the 23 projects, the coding strategy was kept as straightforward as possible, with a focus on identifying the most important themes, and similarities and differences across the case study categories. The key themes that emerged are analysed and discussed below.

Since the meta-evaluation project was undertaken and funded under CECAN, it was constituted and administered as a CECAN research project. The design of the project, implementation, findings and final reporting of the project were reviewed at key stages by two independent senior academic members of CECAN (external to CEP), who were established as a Steering Group for the project, and met with the project team in person. This provided important external validation by experienced evaluators of the findings by the internal team. The findings were, therefore, subject to peer review.

Findings

Describing the Data

Summary descriptive statistics[4] are provided here relating to the sample evaluations and the research questions. Specifically, we describe the characterisations of the projects in terms of:

1. **Geographical scale and policy area**. The left side of Figure 1 shows the range of geographical scale of the projects, from EU level through to the local level and the right side highlights that all policy areas examined were nexus issues, within the natural

[4] Note that some projects meet multiple criteria and so total numbers in Figures 1 and 2 can sum to greater than n = 23.

environment policy area. Example nexus issues included: land use, biodiversity offsetting, rural development, nature improvement areas, EU cohesion policy on environment. All of these areas require interdisciplinary approaches together with inclusion of key stakeholders.

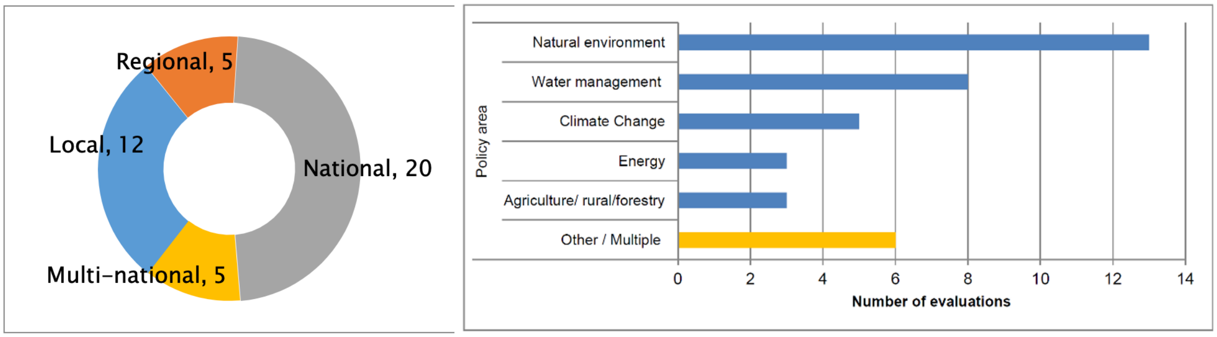

Figure 1. *Geographical scale (left) and range of policy areas covered (right).*

2. **Types of evaluation carried out**. Formative and summative evaluations were found to be commonly used together to satisfy project needs (12 projects satisfied both categorisations), and all of the developmental evaluations were also formative (5 out of 5). Almost half of all participatory evaluations were quasi-experimental (5 out of 11) and all of the ex-ante evaluations were quasi experimental. None of the ex-post evaluations included an ex-ante evaluation of the project. Overall, it is clear that a range of evaluation approaches are needed to tackle these complex areas.

3. **Types of complexity found in the projects**. The top five most common types of complexity found in the projects (see Figure 2) were:

- Policy/Response-related complexity:
  - Multiple agencies/ actors/ stakeholders involved or targeted by the policy (may include conflicting interests)
  - High degree of flexibility or tailoring/changes in the policy during implementation
- Problem-related complexity:
  - Problem has multiple elements
- Impact-related complexity:
  - Poor availability of information and monitoring data relating to impacts
  - Lack of clarity in the causality between actions and impacts (difficulty in attributing causality)

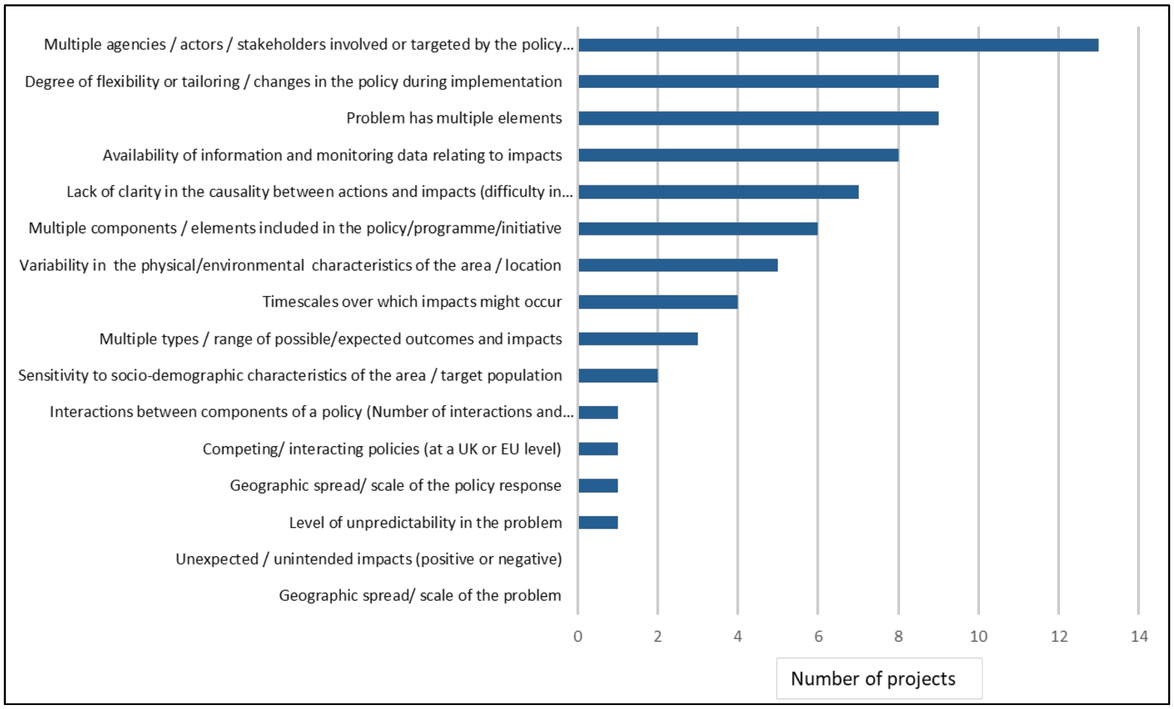

Figure 2. *Types of complexity found in the projects [the top three complexity issues were noted for each project; totals add to >23; refer also to Table 1].*

While some issues around complexity were faced by all types of projects, such as the availability of information and monitoring data, others tended to be specific to the nature of the project. As such 5 out of 7 projects dealing with a lack of clarity in the causality chain between actions and impacts were Programme level initiatives as were the majority of projects (5/7) dealing with multiple components.

Policy/response-related complexity is the main source of complexity existing in EU policy interventions, while it remains a considerable portion of identified complexity in Programme level initiatives.

4. **Types of evaluation use found in the projects.** All of the EU policy intervention evaluations had instrumental use (see Figure 3). The National Policy intervention evaluations also tended to have instrumental and conceptual use, though none of them had process-related use. Programme level policies and initiatives were the only ones that had process-related use. Only a very small number of projects (2 out of 23) covered all 4 types of evaluation use and almost half of the evaluations that had strategic use also had process related use (4/10) and almost all of those (3/4) also had conceptual use.

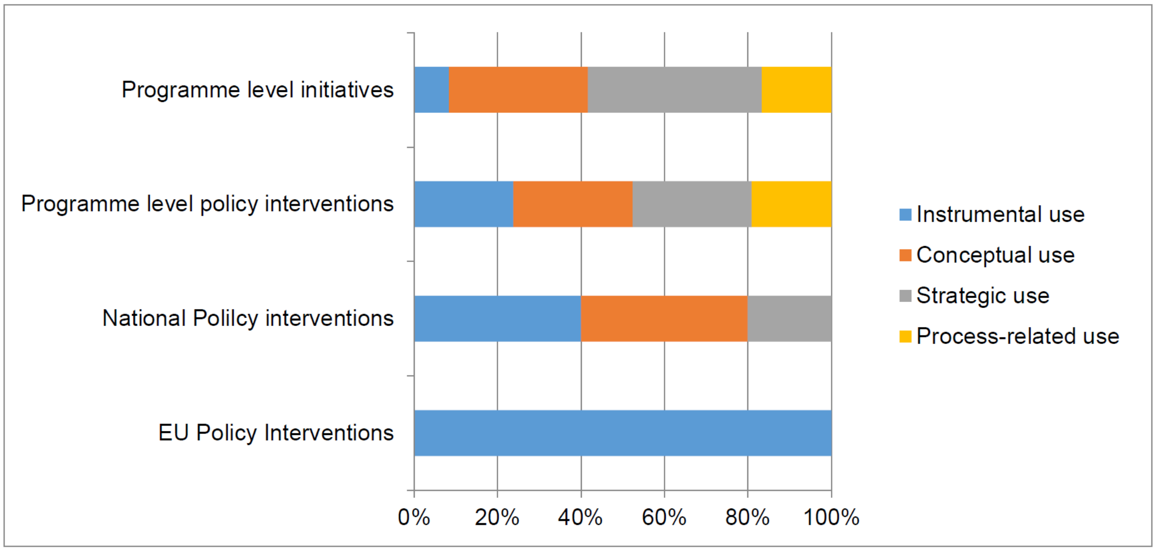

Figure 3. *Types of evaluation use exhibited by case category.*

Analysis and Discussion: Answering the Research Questions

The four overarching research questions provide the focus for the analysis and meta-evaluation.

Were the Evaluations Fit for Purpose, and was Their Purpose Clear? What Lessons can we Learn About Assessing the Effectiveness of the Policy Interventions?

The main distinction found in terms of fitness for purpose was between UK and EU evaluations, and in particular UK programme level initiatives. The evaluations happening in the absence of a clear policy context—among those projects CEP had evaluated—appear less likely to have an explicit theory of change already articulated. Some of these more exploratory interventions have learning and process as defining characteristics and require more attention to dialogue among stakeholders to avoid inconsistent evaluation objectives. Even policy-level interventions may lack an explicit theory of change and/or unclear objectives. In general, the evaluations were fit for purpose inasmuch as they ended up often being tailor-made because of the evolving policy context and the need for flexibility in establishing and modifying evaluation objectives. But evaluation cannot substitute for a strong policy process or clear policy purpose; it can question the policy/policy intervention, but it is only one element among many that determines policy. Two key themes emerged around evaluating effectiveness:

1. **Clarity and consensus of objectives of the policy/intervention and of the evaluation.** Interventions of an exploratory nature, where learning and process are the defining characteristics, seem to represent ambiguity to the pathway of implementation. This requires more attention from stakeholders to avoid vaguely defined/inconsistent policy and evaluation objectives. It is important to ensure objectives of a policy/intervention are linked to a clear baseline and that there are specific measurable outcomes that an evaluation can then assess. Scope to discuss, amend and agree evaluation objectives as part of the initial work on an evaluation helps ensure clarity and fitness for purpose, and ongoing reflection on evaluation objectives is important especially when the policy objectives may be evolving over the time of the project. Where projects are

linked in a wider programme, then setting clear programme level objectives at the outset to reflect the relationship between the programme and project level can aid robust evaluation. Full impact evaluation may not be possible for some complex policy interventions, especially where these are delivered over relatively short timescales. Scoping early in the policy design phase of what is possible for an evaluation to deliver would be helpful.

2. **Stability of governance/policy context and the role of effective project management**. Complex policy interventions often require the involvement of diverse stakeholder groups, which means that different expectations, roles and views on objectives and progress will need to be considered and time needs to be allocated to getting agreement on objectives and evaluation. Time is required to develop a good working relationship with the commissioning project manager to ensure that any issues around contrasting views on project boards are managed. Time available may be affected by tight project timeframes.

How has the Framing of the Evaluation Been More or Less Useful for Understanding Complexity?

What was clear from the analysis was that no one framework was used exclusively across the CEP evaluation projects. Each project fitted into between 2 and 5 types of evaluation categories. This stems in part from the origins of CEP's evaluation work, which comes out of having expertise in nexus topics rather than being solely evaluation experts, and in part because of the type of evaluation requested by clients. Overall, the use of logic models has been widespread in the CEP sample and generally more explicit in recent years with the emphasis on the Magenta book being specified in tenders. Policies, however, are often lacking an explicit theory of change and the evaluation may be the first time such a theory of change has been articulated. Long term impacts, for example, in relation to biodiversity or flooding, are not capable of being evaluated within typical timescales for evaluations (2-3 years). Therefore, an emphasis on *outcomes* as the focus becomes necessary alongside a theory of change to understand how outcomes relate to long term intended impacts. Two key themes emerged around the use of framing of evaluations for understanding complexity:

1. **Timescales**. In designing an evaluation, it is important to recognise that timescales of delivery (activities and outputs) may differ from intervention outcomes and impacts, and that many impacts, especially in natural environment initiatives, cannot be detected over time periods of less than 5 years and in some cases decades. Where possible, therefore, longer-term monitoring should build on existing data and plan for the re-assessment of key indicators after the funded intervention has completed.

2. **Frameworks**. An effective evaluation is likely to require an evaluation framework supported by, for example, a clear logic model. Given the potential for delays between activities and outcomes and impacts a theory of change model(s) is a useful approach, accompanied by mechanisms for testing/validating the theory of change.

Types of Methods for Types of Complexity?

What methods have been used for dealing with aspects of complexity found within environmental policy? Which methods appear to have been most effective? Were some methods and techniques more suited to certain types of complexity? All projects used a mixed-methods approach to data gathering, for example, documents, interviews, surveys, etc. and frequently both quantitative and qualitative data. Qualitative data (collected through interviews, expert advice, workshops) were used more frequently than quantitative data. Qualitative data focuses on description, explanation and in understanding the context in which impacts might be realised. A mixed-method approach allows triangulation of data and helps capture the perspectives of different stakeholders in different depths as necessary. Further, a mixed approach can allow consistent monitoring and evaluation for some objectives and more flexible reporting to reflect local objectives.

The top five types of complexity (from across the three categories of complexity) identified across the projects were:

- Policy/Response-related complexity:
  - Multiple agencies/actors/stakeholders involved or targeted by the policy (may include conflicting interests)
  - High degree of flexibility or tailoring/changes in the policy during implementation
- Problem-related complexity:
  - Problem has multiple elements
- Impact-related complexity:
  - Poor availability of information and monitoring data relating to impacts
  - Lack of clarity in the causality between actions and impacts (difficulty in attributing causality)

Looking across the four case categories, some observations about the relationship between methods and types of complexity can be made. Across the European Union (EU) projects, the most commonly seen types of complexity were: multiple agencies/ stakeholders involved; flexible implementation (e.g., between EU and Member State (MS) levels); and availability of data/indicators. The most common methods used were: interviews, surveys and steering groups/expert advice. This suggests that stakeholder-led methods may have been used to help address complexity in implementation, stakeholder numbers/diversity and where there are limited data/indicators.

For the National Policy intervention (NP) projects there was limited evidence of an association between methods used and types of complexity. However, all three evaluations reviewed made use of interviews and were also characterised by complexity related to the availability of evidence/data related to impacts. The use of interviews (and surveys etc.) is a method that enables perceptions of change or impact to be gathered and assessed in the absence of data/indicators.

More so than in any other category of evaluations, the evaluation of Programme level Policy Intervention (PPI) projects involved undertaking a literature review, and using steering groups or groups of experts to collect evidence. Surveys and observational data were rarely used, while workshops were more common than usual along with interviews. The latter as a choice of evidence collection methods, links to the identification of 'Multiple agencies/actors/stakeholders involved or targeted by the policy', as the most commonly identified complexity criterion in the PPI case category.

Finally, for the Programme level Initiatives (PI) projects the available information doesn't provide a clear link between the methods used and the complexities indicated across the projects. However, the most common types of complexity are characterised by 'multiplicity of factors': e.g., multiple agents/actors; problem has multiple elements; multiple components included in the initiatives. Another commonly identified complexity was the lack of clarity in the causality between actions and impacts. The most common methods used were: interviews and surveys, and the use of causal chain analysis for mapping causality pathways. This could indicate that in order to deal with the variety of actors/elements etc. engaging stakeholders was considered to be the best way forward.

Four key themes emerged in relation to the appropriateness of evaluation methods:

1. Types of methods
   Qualitative and mixed methods are well-suited to addressing complexity in nexus-related evaluations.

2. Data
   The use of existing national datasets, and centralised analysis, where possible can help support effective, robust and efficient evaluation at both programme and local levels. Self-reported data and locally specific indicators can play a useful role; however, such approaches require support and facilitation, and therefore resources, and may result in inconsistent data.

3. Resources
   Careful consideration is needed in the commissioning and design of bespoke information technology (IT) systems for

short-term policy interventions to ensure that they are proportionate and provide value for money, taking into account the design, maintenance implementation and support costs.

4. **Policy development**
   Explicit options appraisal in policy development (ex-ante assessment) can help inform counterfactual analysis (ex post), providing clear linkage between the different types of assessment/evaluation.

What Factors Lead to an Evaluation Being More (or Less) Influential in Policy Changes/Outcomes/Evaluation Use?

The existence of a strong or weak policy cycle and stable/evolving policy appears critical if evaluation is to have instrumental use, i.e. the evaluation needs to have somewhere to go—to feed into. This is what occurs in the typical EU policy cycle, where evaluations are frequently part of a formal and structured review process of legislative instruments, principally for accountability (Schoenefeld & Jordan, 2019). Otherwise the extent to which the evaluation has any influence is dependent on more arbitrary factors, for example, the interest of a minister in a particular policy; change of policy priorities etc. and subject to the vagaries of an evolving policy in flux, under a system stewardship-type model.

Overall, the analysis showed a low level of instrumental use of evaluation in UK programme initiatives and policy interventions and a high proportion of strategic use. That does not mean that evaluations are not being used, just that strategic use—for accountability and defending/promoting policy—may imply that evaluations were used, where they provide the appropriate answers, to support policy development, or where they do not may be used as part of the rationale for dropping a certain policy direction or intervention (though it may actually have been for a range of other political or budgetary expediency purposes). At least two policy interventions in the meta-evaluation hit the buffers as policy interventions—because of lack of funding or because it became a political hot potato/non-starter. In such cases the evaluations were also equivocal—at best they were lukewarm, identifying only marginal benefits and in the case of one considerable costs and risks. In both cases policy was highly fluid—examples of 'system stewardship' perhaps (rather than a systematic policy cycle).

Two key themes emerged in relation to the policy use or impact of evaluation studies:

1. **Nature of the policy process**
   High level of instrumental use is seen in EU policy evaluations, because they are designed to deliver that within a strong policy cycle. Much of UK environmental policy-making exhibits a high degree of flux—more typical of a system stewardship model of policy making/governance than a typical policy cycle. Consequently, evaluation has to be nimbler and more flexible to respond to ongoing changes in policy purpose, design and implementation. Evaluation can have influence in more indirect ways (see Peck and Gorzalski, 2009)—conceptual, strategic or process influence—and these appear more likely in a system stewardship model of policy making.

2. **Knowledge about how evaluations were used**
   An important human factor that influenced an assessment of evaluation use was minimal post-evaluation interaction with evaluators, due to the contractual nature of the projects reviewed in this study.

Limitations of the Meta-Evaluation

This was clearly not a random sample of evaluations, reflecting the nature, focus and types of commissioned evaluations for which CEP has expertise, for which it bids and is successful in securing. But the multiple, embedded case study approach, nonetheless, allows for valuable insights into the range and nature of evaluations undertaken by the case study company. A meta-evaluation of this nature, undertaken by one company, is a rare occurrence. Rarely do commercial consultancies have the opportunity, as we did under the auspices of CECAN[5], to undertake such a retrospective review of their own projects to learn lessons. It would be interesting to compare CEP's experience to that of other evaluation consultancies in the same and different policy sectors. By their nature, the experience of individual contractors will be unique to their own expertise and skill set. These are not factors that can be controlled for, but are important contextual factors that shape the way complex evaluations are approached and delivered.

Some caution is needed in drawing and relating conclusions from this study too broadly, given the nature of the company involved, the selected projects evaluated, and the fact that the meta-evaluation was undertaken internally of projects originally evaluated by the same team. However, while the meta-evaluation was undertaken internally, it was nevertheless subject to independent external guidance, review and validation undertaken as it was as a CECAN research project. This provides a greater degree of confidence in the validity and applicability of the findings to future evaluations in similar nexus areas.

Conclusions

Government often tries out policy ideas, for example, through piloting, but financial support for new policy interventions is often short term (perhaps 2-3 years), or the policy shifts with a change of priority/minister. This has significant implications for evaluation since the model of evaluation in many people's head is as a part of a policy cycle, which evaluates implementation and feeds back into revision of purpose and objectives of policy. In a flux state ('system stewardship') evaluation can feed into purpose, implementation and design all at the same time. It means that the use or influence of evaluation needs to be considered in much broader terms than simply direct or instrumental use, extending evaluation's role into conceptual, strategic and process use/influence/impact, which may be much harder to unravel. Direct/instrumental use was found in the CEP meta-evaluation to be more typical of rigid policy and evaluation frameworks found in the context of EU policy/Directives, in contrast to conceptual, strategic or process use in UK evaluations. These also reflected the different natures and purpose of the policy interventions being evaluated.

In many of the UK cases an explicit theory of change was first elaborated only as part of the evaluation process, after the policy intervention had been initiated and sometimes after it had been running for some considerable time (months or years). An important lesson from this for policy making more generally is the need for policy to be more explicit as to its objectives and intervention logic—what is it trying to achieve and how is it expected to achieve it? Especially in a dynamic system stewardship model—where there is iteration among the purpose, design and implementation of policy—being clear about the purpose is essential and having a theory of change from inception means there is a theory that can be validated, modified and revised dynamically as evidence becomes available. A theory of change is needed—under system stewardship or a more traditional policy cycle—precisely so there is clarity when objectives are modified or expectations change as the policy evolves.

The EU evaluations invariably have more rigid prescription regarding monitoring, indicators, and evaluation frameworks and evaluation questions because of the need for consistency and comparability across all EU Member States, including the use of mixed methods (especially the use of formalised regular reporting and quantitative indicators) and use of qualitative semi-structured interviews/focus groups with stakeholders. The evaluations invariably have instrumental use at the EU level because they are designed to do just that, in comparison to many UK evaluations which appear from this meta-evaluation (inasmuch as there is evidence that they are actually used) to have more strategic or conceptual use as they feed in a more dynamic way into policy evolution.

[5] Centre for the Evaluation of Complexity Across the Nexus, of which CEP was a founding member (see www.cecan.ac.uk).

A series of questions emerges from the findings above, that practitioners (and indeed commissioners) could usefully ask themselves when starting out on a new complex policy evaluation.

Key questions for new evaluations:

- What is the nature of the policy context in which your evaluation is being carried out? Would you describe it as evolving, stable, unclear, high profile?
- How far are the objectives of the *policy* or intervention/initiative clear and amenable to evaluation? Are the expected outcomes and impacts clear?
- How far are the objectives of the *evaluation* clear and achievable given the nature/timing of the policy/intervention/initiative and the resources of the evaluation?
- Are there multiple stakeholders involved as part of the steering group for the policy intervention/initiative? How far is there consensus across perspectives? Are there clear mechanisms in place to enable management of different perspectives?
- Is there a clear and active Project Manager in the policy institution for the evaluation?
- What are the expectations of the client in relation to the ability of the evaluation to evaluate longer term impacts?
- What types of complexity are most relevant to the evaluation?
- To what extent do you think your methods are appropriate for evaluating these complexities? What strategies can you use to address these specific aspects of complexity?
- What types of uses or impacts are expected by your evaluation? How will the client assess whether they have been realised?
- How can you improve the uses or impacts of your evaluation? Where are the points of influence within the evaluation?

A better understanding of the policy context, its state of flux or stability as well as clarity of policy intervention's objectives (and theory of change) could help significantly in designing policy evaluations that can deliver real value for policy makers. Recognising that instrumental use is not necessarily the 'gold-standard' for evaluation impact is also important; evaluations have other valuable uses and can be tailored to maximise those values where such potential impact is recognised.

Acknowledgments

The authors acknowledge the support of The Centre for the Evaluation of Complexity Across the Nexus (CECAN), with funding from UK Research and Innovation/Economic and Social Research Council (UKRI/ESRC).

References

Balthasar, A. (2009) 'Institutional Design and Utilization of Evaluation: A Contribution to a Theory of Evaluation Influence Based on Swiss Experience', *Evaluation Review*, 33(3), pp. 226–256. doi: 10.1177/0193841X08322068.

BetterEvaluation (no date) Report and support use. Available at: https://www.betterevaluation.org/en/rainbow_framework/report_support_use, accessed 17 June 2020..

CECAN (2018) Policy evaluation for a complex world. April 2018, Version 2.0 online at www.cecan.ac.uk, accessed 17 June 2020.

Downe, J., Martin, S. and Bovaird, T. (2012) 'Learning from complex policy evaluations', *Policy and Politics*, 40(4), pp. 505–523. doi: 10.1332/030557312X645766.

European Environment Agency (2011) Bridging long-term scenario and strategy analysis- organisation and methods, Technical Report No 5/2011. doi: 10.2800/76903.

Glouberman, S., & Zimmerman, B. (2002) 'Complicated and complex systems: what would successful reform of Medicare look like?', *Romanow Papers*, 2, pp. 21–53.

Goodin, R. E., Moran, M. and Rein, M. (2008) *The Oxford Handbook of Public Policy*. Oxford: OUP.

Hallsworth, M. (2011) 'System Stewardship' The future of policy making? Working

Paper, London: Institute for Government, pp. 1–49. Available at: https://www.instituteforgovernment.org.uk/sites/default/files/publications/System%20Stewardship.pdf, accessed 17 June 2020.

Hargreaves, M. B. and Podems, D. (2012) 'Advancing Systems Thinking in Evaluation: A Review of Four Publications', *American Journal of Evaluation*, 33(3), pp. 462–470.

Henry, G. T. and Mark, M. M. (2016) 'Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions', *American Journal of Evaluation*, 24(3), pp. 293–314.

HM Treasury (2018) *The Green Book: Central Government guidance on appraisal and evaluation.* London: HM Treasury. Available at https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-governent, accessed 17 June 2020.

HM Treasury (2020a) *The Magenta Book: Central Government guidance on evaluation*, March 2020, available at https://www.gov.uk/government/publications/the-magenta-book, accessed 17 June 2020.

HM Treasury (2020b) *The Magenta Book: Supplementary Guide: Handling Complexity in Policy Evaluation*, March 2020, available at https://www.gov.uk/government/publications/the-magenta-book, accessed 17 June 2020.

Jaffe, A. B., Newell, R. G., & Stavins, R. N. (2005) 'A tale of two market failures: Technology and environmental policy', *Ecological Economics*, 54(2), pp. 164-174.

Kirkhart, K. (2000) 'Reconceptualizing evaluation use: An integrated theory of influence.', in Caracelli, V. and Preskill, H. (eds) *The expanding scope of evaluation use*. New Direct. San Francisco, CA: Jossey-Bass., pp. 5–24.

Magro, E. and Wilson, J. R. (2013) 'Complex innovation policy systems: Towards an evaluation mix', *Research Policy*, 42(9), pp. 1647–1656.

Patton, M. Q. (2002) *Qualitative Research and Evaluation Methods*. 3rd edn. London: Sage Publications.

Pawson, R. et al. (2005) 'Realist review--a new method of systematic review designed for complex policy interventions.', *Journal of Health Services Research & Policy*, 10 Suppl 1(July), pp. 21–34.

Pawson, R. (2013) *The science of evaluation: a realist manifesto*. Sage.

Peck, L. R. and Gorzalski, L. M. (2009) 'An Evaluation Use Framework and Empirical Assessment', *Journal of MultiDisciplinary Evaluation*, 6 (12), pp. 139–156.

Rogers, P. J. (2008) 'Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions', *Evaluation*, 14(1), pp. 29–48. doi: 10.1177/1356389007084674.

Sanderson, I. (2000) 'Evaluation in complex policy systems.', *Evaluation*, 6(4), pp. 433–454.

Schoenefeld, J. J. & Jordan, A. J. (2019) Environmental policy evaluation in the EU: between learning, accountability, and political opportunities? *Environmental Politics,* 28:2, 365-384.

Snyder, S. (2013), "The Simple, the Complicated, and the Complex: Educational Reform Through the Lens of Complexity Theory", *OECD Education Working Papers*, No. 96, OECD Publishing.

Warburton, D., Wilson, R. and Rainbow, E. (2010) 'Making a Difference: A guide to Evaluating Public Participation in Central Government', pp. 1–47. Available at: https://www.involve.org.uk/sites/default/files/uploads/Making-a-Difference-.pdf accessed 17 June 2020.

Wong, G. et al. (2014) 'Development of methodological guidance, publication standards and training materials for realist and meta-narrative reviews: the RAMESES (Realist And Meta-narrative Evidence Syntheses – Evolving Standards) project', *Health Services and Delivery Research*, 2(30), pp. 1–252. doi: 10.3310/hsdr02300.