Environmental Science and Policy 116 (2021) 258–265

Contents lists available at ScienceDirect

Environmental Science and Policy

journal homepage: www.elsevier.com/locate/envsci

A synthesis of the frameworks available to guide evaluations of research impact at the interface of environmental science, policy and practice

Elena Louder [a,*], Carina Wyborn [b], Christopher Cvitanovic [c,d], Angela T. Bednarek [e]

[a] School of Geography, Development, and Environment, University of Arizona, Tucson, USA
[b] Institute for Water Futures, Australian National University, Canberra, Australia
[c] Australian National Centre for the Public Awareness of Science, Australian National University, Canberra, Australia
[d] Centre for Marine Socioecology, University of Tasmania, Australia
[e] The Pew Charitable Trusts, Washington, DC, USA

ARTICLE INFO

Keywords:
Science-policy
Evaluation
Knowledge exchange
Co-production
Boundary-spanning

ABSTRACT

Evaluating the impacts of environmental science on policy and practice is inherently challenging. Impacts can take a variety of forms, occur over protracted timeframes and often involve subtle and hard-to-track changes. As a result, diverse impacts are impossible to capture through traditional academic metrics such as publications and citations, and cannot be captured by focusing solely on end results of a given research project, such as changes in policy or practice. However, despite these challenges, environmental scientists are increasingly required to demonstrate the impact of their work, for example, in funding applications or for career progression. As a result, there has been increased effort among academics and practitioners alike to develop frameworks to guide the evaluation of impacts at the intersection of environmental science, policy, and practice. In this paper we synthesize this rapidly developing landscape of evaluation frameworks. Drawing from literature across fields such as co-production, knowledge exchange, boundary-spanning and other related subdisciplines, we explore common themes and areas of divergence across the different evaluation frameworks. Through qualitative analysis we show that the differences between frameworks often trace back how knowledge is understood and what counts as impact. We conclude by reflecting on our analysis, and articulating 'rules of thumb' to help guide the selection of an evaluation framework. In doing so, we hope that this synthesis contributes towards a growing community of practice aimed at supporting an improved relationship between environmental science, policy and practice.

1. Introduction

Academics, practitioners, and research funders increasingly seek to understand and evaluate research impacts. Especially in the context of complex environmental challenges, researchers are expected to advance scientific knowledge *and* demonstrably contribute to addressing societal, environmental, or economic problems (Lubchenco, 1998). However, evaluating the impacts of environmental science on policy and practice is particularly challenging given that the relationship between research activity and real-world impact is often nonlinear, difficult to trace, and occurs over mis-matched time-scales (Pitt et al., 2018; Posner and Cvitanovic, 2019; Posner et al., 2020). Further, directly attributing impact to a particular research activity is nearly impossible, as policy and practice decisions are informed by diverse values and multiple sources of knowledge, and are shaped by power dynamics beyond the direct influence of research activities (Hakkarainen et al., 2020).

Underlying challenges to measuring impact are different understandings of what exactly impact is and looks like. For example, a number of scholars describe different ways that impacts can be characterized, including instrumental impacts, conceptual impacts, attitudinal impacts, cultural impacts, and/or impacts related to capacity building and preparedness (Reed et al., 2018; Fazey et al., 2014; Wyborn et al., 2018; Edwards and Meagher, 2019). This matters for evaluation because impact that is considered to be direct changes in policy or practice (i.e. new legislation) will require very different frameworks than impacts that are more conceptual (i.e. new framings or understandings). Further complicating matters, conceptual impacts may be precursors to more obvious changes in policy or practice.

Overcoming these challenges and understanding impact, however, can improve accountability to all involved. Research funders,

* Corresponding author.
E-mail address: elouder@email.arizona.edu (E. Louder).

https://doi.org/10.1016/j.envsci.2020.12.006
Received 28 July 2020; Received in revised form 3 December 2020; Accepted 13 December 2020
Available online 30 December 2020

stakeholders, research organizations, and researchers themselves need evidence that efforts to influence policy and practice are an effective use of time and resources, and actually contribute to their respective goals (Henrick et al., 2017; Pitt et al., 2018). Since efforts to influence policy and practice are often misunderstood, not clearly defined, or invisible, effective impact evaluation can also demonstrate the benefits of this work, bolstering the reputations of researchers working in this space, and provide essential evidence of effectiveness (Edwards and Meagher, 2019; Hansson and Polk, 2018; Maag et al., 2018, Bednark et al., 2018). Not least of all, understanding impact can further advance boundary spanning efforts (Edwards and Meagher, 2019), and improve our ability to assess impact and promote effective practice (Henrick et al., 2017).

In light of this, there is increased attention towards improving the evaluation of impacts of environmental science on policy and practice through integrative modes of research and engagement. Here we take integrative modes of engagement to encompass the various ways people and organizations operate at the interface of science, policy, and practice, encompassed by notions of co-production (Nortström et al., 2020), knowledge exchange (Cvitanovic et al., 2015), boundary–spanning (Bednarek et al., 2018), actionable research (Arnott et al., 2020), transdisciplinarity (Hansson and Polk, 2018) and similar. These efforts focus on supporting sustainability outcomes through a variety of approaches to interactive knowledge development and have, in turn, led to an explosion of evaluation frameworks which go beyond traditional academic metrics of impact. These diverse frameworks share common goals: they attempt to capture the diverse types of impacts emerging from alternative ways of working at the science policy interface, to illustrate the validity and utility of these approaches, and to build on understandings of the effectiveness to improve practice.

These frameworks, however, emerge from different domains and disciplines, use differing terminology, and approach evaluation from different founding assumptions (Pitt et al., 2018; Leith et al., 2018); . Thus, in this rapidly developing field, it can be hard to make sense of the various frameworks, and it is confusing for funders, researchers, and practitioners to know what works in what contexts and why. Given the overlaps and similarities, there may be missed opportunity for learning from shared experiences, challenges, and successes (Posner and Cvitanovic, 2019). Therefore, we contend that there is a need to reflect on the frameworks currently used to guide research impact evaluations, so as to improve the practice and implementation of integrative modes of research and engagement.

We seek to address this gap by undertaking a synthesis of frameworks used to guide the evaluation of impacts at the interface of environmental science, policy, and practice. Specifically, we examine the epistemological foundations and assumptions of these frameworks, draw out their similarities and differences, and collate common findings to guide future research and continue to improve the evaluation and practice of co-production and boundary spanning. We start by explaining out approach and methods, which comprises a qualitative literature review of academic articles and grey literature, and a qualitative analysis of evaluation frameworks and indicators from a subset of these materials. We then present the results of our analysis, followed by a discussion reflecting on the similarities, differences, points of overlap and divergence in the frameworks. The paper concludes by offering guidance for selecting frameworks and indicators to guide evaluations at the science and policy interface.

Finally, we recognize that this field is rapidly evolving, and that our selection of emerging frameworks may not be comprehensive. However, this paper presents an early attempt to draw common lessons from across the different frameworks and contribute a novel meta-level synthesis. Furthermore, this article is an effort to contribute to a community of practice oriented toward learning and improving boundary spanning and co-production (Posner and Cvitanovic, 2019; Posner et al., 2020; Pitt et al., 2018).

2. Methods

To address the aims of this study we approached our analysis in two parts. First, we conducted a review of published peer-reviewed and grey literature that presents a framework for evaluating the impact of environmental science on policy and practice. In the second part, we conducted a qualitative analysis of a subset of these frameworks which offered specific indicators to guide impact evaluation.

2.1. Literature review

The literature review consisted of 27 peer-reviewed articles and six grey literature sources from across the subfields of knowledge exchange, knowledge brokerage, boundary spanning, transdisciplinary research, responsible research for innovation, reconciling research supply and demand, and general research impact, and were not limited to environmental literature. They were identified based on the experiential knowledge of the research team who are active in these research fields, and supplemented through advice provided by other disciplinary experts following a call on Twitter. We were particularly interested in frameworks that are currently being used in the field; this motivated our decision to select literature based on the expertise of the authors and other experts (via Twitter) rather than using databases and search terms. The six grey literature sources - The WT Grant Foundation, The Wellcome Trust, Linking Knowledge with Action from the Packard Foundation, Research Quality Plus from IDRC, The Australian Centre for International Agricultural Research RAPID Assessment, and the London School of Economics Impact Blog - were selected based on their reputation for successfully working at the interface of science, policy, and practice. The literature review process consisted of reading all papers in their entirety. We began this process by examining review papers which presented general information on how to approach an evaluation and how to think through the selection of evaluation methodologies. We then moved on to literature which presented more specific guidelines or recommendations on substantive principles and attributes to be evaluated. In this review process, recurrent themes were identified and compared across the literature to highlight points of similarity and divergence. The themes and concepts identified in the literature review were then used to develop the initial coding structure of our qualitative analysis (see below).

2.2. Qualitative analysis

The analysis of frameworks consisted of 14 sources: eight peer-reviewed publications and six grey literature sources (two websites, three reports, and one blog), selected from the literature review described above. The inclusion criteria for the qualitative analysis was the presence of actual indicators, questions, or rubric criteria to guide impact evaluations. This allowed us to eliminate many papers/sources which provide just broad principles, and focus on specific and tangible ways impact is evaluated. The two exceptions are a blog post from Holmes, 2017Holmes (2017) at the London School of Economics and Political Science, and a webpage from the Wellcome Trust which do not contain indicators, but we included because their principles were of particular interest as they come from leaders working at the intersection of science, policy, and practice.

Our analysis began through an open coding process (Strauss and Corbin, 1998) where we sorted the more abstract elements of the frameworks, frequently referred to as principles, into conceptually related parent level codes. The initial codes were drawn from the broader literature review, and included things like: research design, project management, outputs, and impacts. We then assembled all the more specific elements of the frameworks (variably referred to as indicators, rubric questions, criteria, or questionnaire content) into related subsidiary codes of the initial parent codes. For example, any indicator asking about things like -which stakeholders are involved in research? is

the problem frame co-designed? is there diverse representation in the researchers? - was coded to 'research design.'

While the initial coding structure was based off themes identified in the literature review, we adapted the coding structure to accommodate our results from the analysis. Following well established coding processes, we started with theory-based codes (identified from the literature) and adapted them to our empirical material. This nonlinear approach – iteratively moving between pre-determined and emerging codes – is commonly accepted in qualitative analysis (Boeiji, 2010). For example, we began with one code for research process, yet after examining the various indicators, divided the code into two distinct codes of 'research process' and 'research management.' Research processes included things like- what were the number and type of research activities? Are there plans for how results will be implemented? whereas 'project management,' included more practical issues like -were findings delivered in a timely manner? was the budget adhered to? In this way, the analysis was iterative, initially informed by themes from the literature, but adapted to the specific principles and indicators from the analyzed materials. This coding structure was then used to identify themes and subthemes which are presented below.

3. Results

3.1. Literature review

Our literature review surfaced a number of papers that offered insights on the logic of designing or selecting an evaluation approach, rather than putting forth a particular evaluation framework. Examples from this category include Gitomer and Crouse (2019); Bornmann (2013); Posner and Cvitanovic (2019); Pitt et al. (2018);Fazey et al., 2014Fazey et al. (2014), and Boaz et al. (2009). In a typical example, Fazey et al. (2014) present a typology of different evaluation approaches, explain the logics behind each, and outline their respective strengths and drawbacks. These articles offer ways to think through the design of an evaluation rather than suggest particular content of an evaluation (i.e. indicators, specific questions, or specific methods of evaluation). For example, De Jong et al., 2011 (p. 62) offer general questions to help guide an evaluation such as, "what are the characteristics of the research field? What is the local context? What are the intended audiences?"

In addition to offering guidelines on how to design an evaluation or select an evaluation approach, the review papers offer considerations for selecting evaluation methods. For example, Posner and Cvitanovic (2019) outline diverse methods of evaluation and what each one captures – for example, interviews can capture stakeholder perspectives; network analysis can measure the presence and strength of relationships between stakeholders; surveys may capture larger sample sizes, yet afford less nuanced responses. Boaz et al. (2009) present a similar pro and con discussion of various methods. They detail how qualitative methods like field observations or semi-structured interviews may capture unexpected and nuanced outcomes, whereas focus groups may be less time consuming and unearth consensus amongst stakeholders. They emphasize that methods must be built around the output of interest. Our analysis revealed a diversity of methods available for this type of work; a complete discussion on methodologies is beyond scope of this paper, however see Posner and Cvitanovic (2019) and Reed et al. (2020) for a more complete review of considerations for choosing methods.

Other articles put forth substantive principles and indicators to guide evaluation. For example, some authors present principles of quality drawing from the credible, salient, legitimate criteria (Belcher et al., 2016; Hansson and Polk, 2018; Posner et al., 2016) and then present indicators within those. Similarly, Wickson and Carew (2014) present the following principles: socially relevant and solution oriented; sustainability and future scanning; diverse and deliberative; reflexive and responsive; rigorous and robust; creative and elegant; honest and accountable. They then present a rubric with ranges of 'quality' from 'exemplary' to 'routine.' Taking a different approach, Maag et al. (2018) present indicators for each stage of the research process (project development to outputs). Whilst others arrange indicators into categories of some variation of 'impact dimension,' for example: context, process, outcome, and impact from Wall et al. (2016), and research problem, research process, and research results from Jahn and Keil (2015).

The reviewed literature also shows a variety of definitions of impact. For some, impact means changes in the world. For example, increased ecosystem health (Posner et al., 2016), societal change (Hansson and Polk, 2018), or changing peoples' lives for the better (Lebel and McLean, 2018). Others break impact down into components or dimensions. For example, Phillipson et al. (2012) break impact into: impacts on the research itself (how did knowledge exchange impact research quality and relevance), and then impacts on the stakeholders (how did knowledge exchange impact stakeholder knowledge and practices). Another frequently cited understanding of impact comes from Nutley et al. (2007) who divide it into: instrumental (direct changes in policy and practice), conceptual (changes in ways of thinking), and capacity building (changes in skills and abilities). Many of the articles are framed in terms of "quality" rather than impact. However, in general, these articles also highlight that a central tenet of transdisciplinarity, (or other subdiscipline) is to contribute to society, the environment, sustainability etc., so drawing a hard line between "impact" and "high quality" could be somewhat artificial.

Yet others offer typologies of impact that encompass many of these domains. For example, Wyborn et al. (2018), outline the following types of impact: conceptual (new ideas, framings or understandings), relationships (new networks and collaborations), strategic (new problem frame), instrumental (new policy or decision), and capacities (new ability to apply knowledge in a given context). In another, Edwards and Meagher (2019) build on the work of Nutley et al. (2007), and offer five dimensions of impact: instrumental, conceptual, capacity-building, enduring connectivity, changes in culture/attitudes toward. Further, Cvitanovic et al. (2018) consider impacts at different levels: impacts on policy and practice; impacts on organizations; and impacts on individuals.

A common theme across the literature is that traditional measures of impact based on academic metrics are insufficient and inappropriate for research in this realm. While quality publications in journals with high impact factors indicate success in traditional research, this body of literature repeatedly suggested that bibliometric indicators do not capture impact at the science policy interface. In one helpful analogy, Leith et al. (2018), describe knowledge co-production activities as an iceberg. The visible tip of the iceberg is the outputs like publications, reports, and other formal knowledge dissemination products. However, they argue, frequently, the most important co-production work is the messy, below the surface, activities where knowledge and viewpoints are wrestled with, ideals and ethics are challenged, and new relationships are built. This analogy harmonizes broadly with other pieces which suggest that formal outputs (like those that would mark the impact of traditional research) tell only a very partial story of the impact of knowledge co-production and boundary spanning.

Another recurrent tension across the literature is the issue of attribution. For example, Posner et al. (2020) suggest that short term outcomes of a project may be clear and easy to attribute, while those which occur over longer time scales may be subtle and challenging to detect. While Posner et al. (2020) emphasize the temporal challenge of attribution, Pitt et al. (2018) highlight how attribution is complicated by multi-stakeholder research projects; where many actors collaborate, tracing a straight line of attribution to a particular individual, organization, or output is challenging. Furthermore, such research often explicitly seeks to foster collaboration, cross-pollination, and collective knowledge construction; attempts to attribute impact to one particular actor can detract from efforts to build shared trust and establish enduring, collaborative relationships (Edwards and Meagher, 2019; Pitt

et al., 2018) In response to this tension, many authors seem to focus more on evaluating the process, or using process indicators as "proxy indicators" of final impact (Meagher and Lyall, 2013).

This literature review demonstrates the great variety of approaches out there. It is therefore unsurprising that there is confusion around the question of what to evaluate, how, and why, when trying to understand the impact of interactive modes of research. A closer look at the detail of the indicators, questions and rubric that is afforded through qualitative analysis can provide more guidance on how to navigate this diversity.

3.2. Qualitative analysis results

Using the coding process described above, both principles and indicators have been grouped into 10 general themes and subthemes. Below is a description of the themes and subthemes, what types of principles and indicators were included in it, and a few examples. The examples given below are a mix of questions, statements, and criteria as that is what was found in the frameworks. In an attempt to paint a picture of the diverse ways people approach evaluation, we have left them as originally formatted.

3.2.1. Research design

This theme contains references to the initial stages of project planning and design. Principles and indicators pertain to the type of questions being asked, the internal logic of the project, and structure of the research. Included here is also the composition of research teams (who participates and how), issues related to methodology, and appropriateness and feasibility of objectives. The subthemes are: project design, objective setting, participation/representation, methodology/research, and expectation management.

3.2.2. Examples

Is project structure suitable for consensus building? (Defila and Di Giulio, 1999); methods are clearly described and documentation demonstrates that the methods are fit to purpose, systematic yet adaptable, and transparent (Belcher et al., 2016); what are the level and type of contributions to the program by research/policy/practice partners? (Maag et al., 2018); scale spanning- consideration of different spatial, temporal, and social scales and of transition effects (Jahn and Keil, 2015).

3.2.3. Context

This theme refers both to factors that are outside of the research project and yet affect it (barriers, politics, and external events) and to the steps taken by research to engage with and understand the context in which research is being carried out. Subthemes include: engagement with context (familiarity and engagement with a situation, place, or actors), barriers, and political context.

3.2.4. Examples

The context is well defined, described, and analyzed sufficiently to identify research entry points (Belcher et al., 2016); constraining and enabling contextual influences- within or external to the research effort-most likely to affect research performance are identified (Ofir et al., 2016); clear and explicit identification of institutional and contextual limitations and a structured effort to acknowledge and improve upon these conditions (Wickson and Carew, 2014).

3.2.5. Enabling conditions for co-production

This theme is related to context but includes references to a broader range of conditions that enable effective research, including incorporation of different types of knowledge and expertise. The subthemes here are multiple stakeholders, multiple epistemologies, expertise, capacity, and funding.

3.2.6. Examples

In-house technical capacity to manage new information (Wall et al., 2016); amount and type of funding, diversity of funding sources, continuing/follow up funding (Maag et al., 2018); necessary scientific disciplines are included on research team (Wall et al., 2016).

3.2.7. Project management

This theme contains references to the ongoing management of research projects, including issues of timing, budgets, and general research execution. Also here is references to working with partners and managing relationships. Also grouped in this theme are principles or indicators related to routines, procedures, or processes put in place to facilitate communication amongst partners. Subthemes include: partnerships, relationships, boundary spanning, and governance.

3.2.8. Examples

Were processes in place to facilitate communication, translation, and mediation? (Davila et al., 2015); are contracts with external participants clear and binding? (Defila and Di Giulio, 1999); are relationships built and maintained? (Maag et al., 2018).

3.2.9. Research process

This theme refers to the stages of the research process where new knowledge is being generated, analyzed, integrated or implemented. Subthemes include: knowledge generation, integration, and implementation.

3.2.10. Examples

The [project] establishes systematic processes for collecting, organizing, analyzing, and synthesizing data (WT Grant Foundation, 2018); number of implementation strategies facilitated (Maag et al., 2018); documentation reflects effective project implementation that is appropriate to context with reflection and adaptation as needed (Belcher et al., 2016).

3.2.11. Attributes of the research

This theme refers to different attributes or qualities of the research. Included in this theme are any references to the salience, legitimacy, and relevance criteria, and references to overall quality and meeting of standards. Subthemes include ethical (e.g transparency, honesty, disclosure), transferable, originality, overall quality, creativity, and output attributes.

3.2.12. Examples

Legitimacy- participants need to be clear on whose behalf they speak (e.g. people in the same profession, users of a particular service, patients with the same condition, employees in a specific organization) and be supported to do so (Holmes, 2017); outcomes work reliably under real-world conditions (Wickson and Carew, 2014); research is carried out to the highest appropriate standards (Wellcome Trust, 2019); integrity, positioning for use, legitimacy, and importance (Ofir et al., 2016).

3.2.13. Communication

This refers to communication within the research process; i.e. internal communication or communication between researchers and stakeholders etc. (as opposed to external communication, or communication of results, which is coded to "outputs"). The subthemes are dialogue, translation, and target audience.

3.2.14. Examples

Frequency and medium of communication between research and management teams (Wall et al., 2016); did investments facilitate flow of information between actors? (Davila et al., 2015); tailor knowledge to needs/contexts of target audiences and transform it into preferred format (Maag et al., 2018).

3.2.15. Outputs

This theme contains references to the end results (products, publications, tools, events) of the research process and their dissemination or reach. Subthemes are tools, products, events, publications, other outputs (all these include reference to presence or production of the products), and dissemination (this contains any references to readership, reach, citations, or quality of the avenues of distribution).

3.2.16. Examples

Results are produced, published, and disseminated (Posner et al., 2016); is there a sufficient number of scientific publications? (Defila and Di Giulio, 1999); number of listeners/viewers, download/visitors, click rate etc.; number of citations/quality of publisher/platform (Maag et al., 2018).

3.2.17. Impacts

This theme contains any reference to changes or impacts that the research aims to or actually does achieve, including impacts resulting from research processes. Subthemes are: social or policy change (changes in society, environment, or policy), build capacity, network building, common ground, change in understanding, and incentives.

3.2.18. Examples

Common language developed, different positions articulated, stakeholder differences mediated (Posner et al., 2016); develop common ground (Maag et al., 2018); findings contribute to climate change adaptation action (Wall et al., 2016); interventions improve the health of many people (Wellcome Trust, 2019); researchers and prospective users stay in contact even after a funded project ends (Meagher and Lyall, 2013).

3.2.19. Evaluation

This theme includes any reference to presence of processes to support evaluation of project processes, outcomes and impact. This node includes any references to learning, reflexivity, reflection, or comparing the project results to its own goals. Subthemes are: reflexivity, learning and meeting goals.

3.2.20. Examples

Number and type of internal documents or publications on lessons learnt (Maag et al., 2018); process of reflection, individually and as a research team, are clearly documented throughout the research process along with clear descriptions and justifications for any changes to the research process made as a result of the reflection (Belcher et al., 2016); structured, periodic review (Wickson and Carew, 2014).

4. Discussion

Here we have presented the results of a literature review and analysis of frameworks used to evaluate the impact of environmental research on policy and practice. In this section, we discuss combined insights from the literature review, as well as the themes and subthemes identified in the qualitative analysis. Our work highlights the diversity of existing frameworks, but also commonalities. We attribute the diversity that is found within the details of the questions, rubrics, and indicators to variation in the design logics, the different approaches to when in the research process an evaluation is intended, and various levels of scope and thoroughness of the frameworks reviewed. Importantly, our analysis surfaced various understandings of impact and divergent underlying understandings of how knowledge is created, both of which have important ramifications for selecting an evaluation approach.

4.1. Various design logics

People arrange principles and indicators according to different design logics, for example, evaluation design according to temporal aspects of a research project, according to of various components of research, or according to a visual model depicting a theory of change. A common approach seems to be to arrange principles roughly corresponding to temporal phases of the research process. For example, Wall et al. (2016) propose: (1) context (2) process (3) output (4) outcomes (5) impacts (6) external factors. Whereas others arrange principles corresponding to "dimensions," (Jahn and Keil, 2015; WT Grant Foundation), "elements," (Wickson and Carew, 2014; Wellcome Trust, 2019), or "attributes," (Belcher et al., 2016). In these examples, "dimensions" and "elements" correspond not to temporal phases of the research process, but to various components of the research that researchers deem important.

4.2. Various approaches to timing

The different logics behind the principles relates to another point of variation in the data: different approaches to timing. Where Maag et al. (2018) propose one evaluation with indicators for each phase of the research to be administered at the end of the research project, Defila and Di Giulio (1999) present three different evaluations to be administered at different points in the research process, and recommend which questions to focus on during which evaluation. Varying still, the Packard Foundation's Linking Knowledge with Action Program (LKwA) initially carried out developmental evaluations with their grantees early in the research process, but discontinued the practice because, they decided, it added an unnecessary burden (Rowe and Lee, 2012). Now LKwA conducts one formative and one summative evaluation. In another approach to timing, Edwards and Meagher (2019) developed an explicitly modular and intentionally flexible framework whereby each component could be applied at any stage of the research process.

4.3. Variations in scope and thoroughness

The data also show great variation in the level of thoroughness and scope. For example, the evaluation from Defila and Di Giulio (1999) contains 16 principles asking a huge array of questions ranging from managerial (was the timeline met? was the budget adhered to? were the managers qualified?) to more substantive questions about research quality (what is the relevance of the overarching project? Are the objectives of the project based in current knowledge?). Others present fewer key principles like Wickson and Carew (2014) who present seven key elements of quality. Decisions about the thoroughness and methodology of an evaluation need to be weighed against the priorities of an organization.

On a more conceptual level, our analysis surfaced various understandings of impact. As Boaz et al. (2009) note, multiple terms are often used interchangeably to describe impact, including outcome, benefit, payback, translation, utilization, etc. Furthermore, our analysis revealed a spectrum, where on one end, impact is only understood as changes in the real world. On the other end, impact includes more intermediate and intangible shifts like changes in knowledge, relationships, or problem framing; for some scholars these changes *are* the impact. To illustrate, the evaluation framework from Posner et al. (2016), presents four pillars of impact pathway, the final pillar being ultimate impact of "improved outcomes for biodiversity, ecosystems health, and human well-being," (p. 1761). In their evaluation, they exclude the final pillar, because changes in ecosystems health etc. cannot be traced back to the research project. Similarly, Hansson and Polk (2018) say that evaluation needs to focus on intermediary effects, such as capacity building and enhanced networks, rather than impact which is defined as transformational and structural changes to society. Here we see scholars suggesting that impact – a thing that happens in the real world- is necessarily beyond the scope of the research, outside the purview of evaluation. In contrast, Edwards and Meagher (2019) put forth a framework where attitudinal and institutional change, and enduring connectivity (between researchers and stakeholders) are

counted as impacts *per se*.

However, what at first seem like serious differences may turn out to be more semantic than substantive. Whether intermediate changes are counted as impact *per se*, contribution to impact, an element of quality, proxy or prelude, many frameworks which put forth a narrow understanding of impact still work to capture more nuanced changes sparked by the research process. One common work-around seems to be by framing evaluation in terms of quality, rather than impact. For example, de Jong et al. (2011) examine quality rather than impact because impact may not yet be visible. Here, scholars again imply that impact is the end-of-pipe changes in the world, and therefore should not be the focus of evaluation. Similarly, Klein (2006) shirk the idea of evaluating impact, opting for quality as a way to measure more holistic benefits of the research process. These two approaches seem to be different ways of describing the same overall phenomenon, and point to an important commonality: whether evaluations expand the meaning of impact, or expand the lens of evaluation to include other dimensions besides impact (defined as changes in the real world) many of the frameworks sought to capture subtle, nuanced, or process-related impacts of research at the science policy interface.

That said, our analysis did surface some notable differences in understandings of impact. Some evaluation approaches conceptualize impact in a relatively simple, end-of-research-process way. For example, Bornmann (2013) says, "society can reap the benefits of successful research studies only if the results are converted into marketable and consumable products (e.g., medicaments, diagnostic tools, machines, and devices) or services," (p.217). Here, impact of research seems to be conceived of only as the beneficial tangible outputs of knowledge generating activities. Others take a much more expansive approach. For example, Posner and Cvitanovic (2019) who outline a wide range of impacts from improved knowledge exchange between scientists and decision makers to more cohesive networks to increased job satisfaction. Still others emphasize the contingent and subjective nature of impact, such as Leith et al. (2018) who argue that, 'people are the outcome.' This phrasing suggests that changes in ways of thinking, acting, or changes in professional networks are clearly 'impacts'. This highlights the fact that where outcomes are often intangible, it is difficult to disentangle the means and processes that lead to positive outcomes from the outcomes themselves (Jahn and Keil, 2015). Where some attempt to maintain a delineation between process and results - for example, Posner et al. (2016: 1763) evaluate process related variables in distinct metrics from knowledge outputs "to maintain a conceptual distinction between knowledge and process," - for others, drawing this sharp line may miss the richness of co-production activities (Jahn and Keil, 2015). Again, this division seems to fall along the lines of those who count changes in understanding and relationships as an impact.

Another important difference is the extent to which evaluations should be based on pre-set expectations. In evaluations based on logic models, evaluation is essentially conceived of as judging outcomes against pre-given expectations (Margoluis, 2009; Nguyen et al., 2017). However, other authors talk about the importance of emergent and contextual characteristics, and say that it's important not to be limited by predefined quality criteria or scientific standards (e.g. Edwards and Meagher, 2019; Hansson and Polk, 2018). Edwards and Meagher (2019) highlight the limitations of the linear causality undergirding logic models that is based on outdated understandings of the way knowledge works. They suggest that the assumption in logic models that one thing leads to another erases the complexity of interactions at the science policy interface. This is not to suggest that that logic models have no use, (indeed many papers emphasized the importance of being clear about assumptions of why something will work), but that it may be important for evaluations to capture emergent, unexpected results in addition to those that were expected.

We trace these differences in evaluation frameworks to the underlying epistemology of a research project. Resonating with Fazey et al. (2014), our review surfaced a spectrum of understandings of knowledge, ranging from more positivist, where knowledge is certain, fixed, and able to be passed along, to more constructivist, where knowledge is conceptualized as always mediated through culture, worldviews, and co-created through various subjectivities. In one poignant example, Leith et al. (2018) discuss how sustainability science is primarily not about delivering new information to decision-makers, rather, it is about opening up and reframing problems and possibilities. In this understanding, work at the science policy interface is not aiming to provide lacking information or knowledge, but to co-create new understandings. The idea of knowledge transfer, or push and pull of knowledge needs and supply is replaced by a more nuanced understanding where knowledge is understood as a social product, filtered through subjective meaning assigned to it, and created through iterative cycles. Put differently, information becomes knowledge only through people's interpretations, meanings, and worldviews (Hakkarainen et al., 2020). These various understandings of knowledge then shape what is meant by impact.

Importantly, we find that how you define impact determines how you measure it. For example, Bornmann (2013) who defines impact as real-world products mentions evaluation methods like patent analysis and scientific citations in policy decisions. The understanding of impact as information-based products necessitates evaluation framework which registers those changes. Conversely, on the other end of the spectrum, understanding of impact as more tacit, subtle changes in understanding, problem framing, and networks requires evaluation indicators which capture those more process- related changes. This is by no means a binary decision, and many frameworks indeed attempt to capture process related changes and product or output related changes. However, answers to underlying philosophical questions like- are people the output, or are people only the means to the output? Is impact in the eye of the beholder, or something objective which can be quantitatively measured? - have implications for selecting an evaluation framework and indicators, and selecting methods.

Finally, we noted a lack of attention to unexpected outcomes. Despite notable exceptions (Edwards and Meagher, 2019; Hakkarainen et al., 2020) we found few detailed discussions on how boundary spanning and co-production can lead to unexpected or even negative outcomes. Hakkarainen et al. (2020) detail the ways in which knowledge was interpreted by stakeholders in ways that were unanticipated by researchers. In their study in Zanzibar, local stakeholders used knowledge instrumentally to change fishing practice, which was not the intention or expectations of researchers. In their case study, Edwards and Meagher (2019) found that one of the key positive impacts from their work was unexpected knock on effects of a carbon flux tower which sparked increased stakeholder engagement. These two examples illustrate the importance of evaluations to capture the broad range of impacts from work at the science policy interface and flag the likelihood of unexpected changes that might occur from research activities. We suggest that this could be a fruitful avenue for future research.

4.4. Rules of thumb for selecting evaluation frameworks and indicators

Based on our review and analysis, we recommend the following rules of thumb to guide the selection of frameworks and indicators for evaluating the impact of research at the interface of environmental science, policy and practice. Given the already overwhelming variety of evaluation frameworks, we offer considerations to bear in mind in selecting among approaches, rather than prescriptions for choosing an evaluation framework.

4.4.1. Be clear about underlying assumptions of knowledge production and definitions of impact

While we do not suggest linear understandings of cause and effect, clarifying from the start how research activities are intended to achieve impact is an important pre-cursor to designing an evaluation. Furthermore, defining what you mean by impact is an important first step in selecting indicators to know if you've achieved it. A research

organization should be clear up front whether changes in attitude, problem framing, or relationships count as impact. For example, if it is assumed that interactions between stakeholders lead to improved relationships, indicators can usefully be developed to evaluate the nature, frequency, quality etc. of interactions. This epistemological clarity helps define what counts as impact, and what counts as robust evidence of that impact.

4.4.2. *Attempt to measure intermediate and process-related impacts*

Whether this means expanding the definition of impact, or evaluating quality, or 'contribution to impact,' select indicators that capture nuanced changes in problem framing, understanding, or mind sets. Our review shows that evaluations should at least partially attempt to capture the 'below the tip of the iceberg' knowledge co-production activities (Leith et al., 2018). This could be done by focusing at least part of an evaluation on measuring perspectives of participants (via interview or survey) regarding changes such as increased capacity, changes in expertise and knowledge, and shifts in how a problem is understood. Attention to intermediate impacts may serve as building blocks for end-of-process outcomes, and also enable the evaluation of 'progress markers' along a theory of change.

4.4.3. *Balance emergent and expected outcomes*

While clear expectations and aspirations are important, evaluations should have at least some open-ended component which captures unexpected outcomes, both positive and negative. As Hakkarainen et al. (2020) suggest, given the dynamism and complexity of work at the science policy interface, outcomes are inevitably unpredictable as various actors create distinct meaning out of shared information. Furthermore, as this field is rapidly developing, the types of impacts that are to be expected is still emerging. This could be implemented through crafting at least part of an evaluation in an open-ended manner. For example, rather than rubrics with pre-determined criteria, Edwards and Meagher (2019) ask instead- what changed? who changed? how do you know? Such an open-ended approach allows for unexpected outcomes to surface.

4.4.4. *Balance indicators that capture nuance and those that simplify*

Evaluations which assign numerical scores to impact may be useful for project managers and large research organizations. However, aggregated scores can overshadow conceptual changes in the way a problem is framed, or subtle changes resulting from knowledge co-production. Overemphasis on simple evaluations can also lead to 'gaming the indicators,' and provide perverse incentives to tailor research to meet the indicators. While indicators that can be quantitatively scored (for a hypothetical example, assigning 1–10 scores on dimensions like suitable context, legitimacy and relevancy, project outputs) may be easy to use, especially for comparing research projects, such an approach might not register *why* or *how* changes occurred. The same is true for the number of indicators- fewer indicators may make evaluation simpler and more convenient, where more indicators may deliver more detailed information. This tension must be considered when designing an evaluation.

5. Conclusion

This paper presents the results of a literature review and qualitative analysis of frameworks to guide the evaluation of the impact of environmental science on policy and practice. Our results show that the frameworks vary in overall design, in scope and thoroughness, in the number of principles and indicators, and the approach to timing and implementation. Importantly, our synthesis suggests that these differences in evaluation frameworks often reflect deeper variation in how knowledge is understood and what counts as impact. However, we note that while some frameworks seek to evaluate broadly defined impact, and others evaluate contribution to impact or quality, a common theme is that evaluation must capture the non-linear, less visible changes in problem framing, mindsets, and relationships. Although these impacts are harder to evidence, work in this area increasingly acknowledges that such changes are an important part of knowledge co-production, and need to be recognized and appreciated. Importantly, there is broad consensus across the literature that traditional metrics of research impact (publications and citations), are insufficient for capturing the diversity and nuance of research at the science policy interface. We have highlighted the important similarities and differences in the extant evaluation frameworks and indicators, and have brought together a novel synthesis in the rapidly evolving field. We hope to have provided insights on how and why different indicators are chosen in different contexts, and contributed to growing calls to develop a community of practice around evaluating impact at the interface of science, policy and practice.

Funding

Support for the project was provided by The Pew Charitable Trusts.

CRediT authorship contribution statement

Elena Louder: Investigation, Writing - original draft. **Carina Wyborn:** Conceptualization, Methodology, Writing - review & editing. **Christopher Cvitanovic:** Conceptualization, Writing - review & editing. **Angela T. Bednarek:** Conceptualization, Writing - review & editing.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

We thank Leo Curran for useful comments on earlier drafts of this manuscript. We also thank all those on Twitter who responded to our call for frameworks for inclusion in our analysis. Finally, we thank the two anonymous reviews who provided constructive comments on an earlier version.

References[1]

Arnott, J.C., Mach, K.J., Wong-Parodi, G., 2020. Editorial overview: the science of actionable knowledge. Curr. Opin. Environ. Sustain. 42, A1–A5.
Bednarek, A.T. et al. (2018). Boundary spanning at the science–policy interface: the practitioners' perspectives. Sustainability Science 13 1175-1183.
*Belcher, B., et al., 2016. Defining research quality in a transdisciplinary context. Res. Eval. 25 (1), 1–17.
Boaz, A., et al., 2009. Assessing the impact of research on policy: a literature review. Sci. Public Policy 36 (4), 255–270.
Boeiji, H., 2010. Analysis in Qualitative Research. Sage Publications, London, UK.
Bornmann, L., 2013. What is a societal impact of research and how can it be assessed? A literature survey. J. Am. Soc. Inf. Sci. 64 (2), 217–233.
Cvitanovic, C., et al., 2015. Improving knowledge exchange among scientists and decision-makers to facilitate the adaptive governance of marine resources: a review of knowledge and research needs. Ocean Coast. Manag. 112, 25–35.
*Cvitanovic, C., et al., 2018. Building university-based boundary organisations that facilitate impacts on environmental policy and practice. PLoS One 13 (9).
*Davila, F., et al., 2015. Knowledge Systems and RAPID Framework for Impact Assessments. Australian Centre for International Agricultural Research.
De Jong, S., et al., 2011. Evaluation of research in context: an approach and two case studies. Res. Eval. 20 (1), 61–72.
*Defila, R., Di Giulio, A., 1999. Evaluating Transdisciplinary Research. *Panorama*: Swiss National Science Foundation News Letter, pp. 4–27.
*Edwards, D.M., Meagher, L.R., 2019. A framework to evaluate the impacts of research on policy and practice: a forestry pilot study. For. Policy Econ. xxx(xxxx):xxx-xxx.
Fazey, I., et al., 2014. Evaluating knowledge exchange in interdisciplinary and multi-stakeholder research. Glob. Environ. Change 25, 204–220.

[1] *denotes papers that were included in the qualitative analysis.

Gitomer, D.H., Crouse, K., 2019. Studying the Use of Research Evidence: A Review of Methods. William T. Grant Foundation.
Hakkarainen, V., et al., 2020. The other end of research: exploring community-level exchanges in small-scale fisheries in Zanzibar. Sustain. Sci. 15 (1), 281–295.
Hansson, S., Polk, M., 2018. Assessing the impact of transdisciplinary research: the usefulness of relevance, credibility, and legitimacy for understanding the link between process and impact. Res. Eval. 27 (2), 32–144.
*Henrick, E.C., et al., 2017. Assessing Research-practice Partnerships: Five Dimensions of Effectiveness. William T. Grant Foundation, New York.
Holmes, B., 2017. On the co-production of research: why we should say what we mean and mean what we say, and learn as we go. London School of Economics Impact Blog. London School of Economics https:blogs.lse.ac.uk/impactofsocialsciences (Accessed 15 August 2019).
*Jahn, T., Keil, F., 2015. An actor-specific guideline for quality assurance in transdisciplinary research. Futures 65, 195–208.
Klein, J., 2006. Afterword: the emergent literature on interdisciplinary and transdisciplinary research evaluation. Res. Eval. 15 (1), 75–80.
Lebel, J., McLean, R., 2018. A better measure of research from the global south. Nature 559, 23–26.
Leith, P., et al., 2018. An operation on 'the neglected heart of science policy': reconciling supply and demand for climate change adaptation research. Environ. Sci. Policy 82, 117–125.
Lubchenco, J., 1998. Entering the century of the environment: a new social contract for science. Science 279, 491–497.
*Maag, S., et al., 2018. Indicators for measuring the contributions of individual knowledge brokers. Environ. Sci. Policy 89, 1–9.
Margoluis, R., et al., 2009. Using conceptual models as a planning and evaluation tool in conservation. Eval. Program Plann. 32 (2), 138–147.
Meagher, L., Lyall, C., 2013. The invisible made visible: using impact evaluations to illuminate and inform the role of knowledge intermediaries. Evid. Policy 9 (3), 409–418.
Nguyen, V., et al., 2017. A roadmap for knowledge exchange mobilization research in conservation and natural resource management. Conserv. Biol. 31 (4), 789–798.
Nortström, A.V., et al., 2020. Principles for knowledge co-production in sustainability research. Nat. Sustain. 3, 182–190.
Nutley, S., et al., 2007. Using Evidence: How Research Can Inform Public Services. Policy Press, Bristol.
*Ofir, Z., et al., 2016. Research Quality Plus: a holistic approach to evaluating research. International Development Research Center (IDRC). Report.
Phillipson, J., et al., 2012. Stakeholder engagement and knowledge exchange in environmental research. J. Environ. Manage. 1 (95), 56–65.
Pitt, R., et al., 2018. Wrestling with the complexity of evaluation for organizations at the boundary of science, policy, and practice. Conserv. Biol. 32 (5), 998–1006.
*Posner, S., Cvitanovic, C., 2019. Evaluating the impacts of boundary spanning activities at the interface of environmental science and policy: a review of progress and future needs. Environ. Sci. Policy 92, 141–151.
Posner, S., et al., 2016. Policy Impact of ecosystem services knowledge. Proc. Natl. Acad. Sci. 113 (7), 1760–1765.
Posner, S., et al., 2020. Boundary spanning among research and policy communities to address the emerging industrial revolution in the ocean. Environ. Sci. Policy 104, 73–81.
Reed, M.S., Bryce, M., Machen, R., 2018. Pathways to policy impact: a new approach for planning and evidencing research impact. Evid. Policy A J. Res. Debate Pract. 14 (3), 431–458.
Reed, M.S., Ferré, M., Martin-Ortega, J., Blanche, R., Lawford-Rolfe, R., Dallimer, M., Holden, J., 2020. Evaluating impact from research: a methodological framework. Res. Policy.
*Rowe, A., Lee, K., 2012. Linking Knowledge with Action: An Approach to Philanthropic Funding of Science for Conservation. A Report to the Conservation and Science Program. David and Lucile Packard Foundation.
Strauss, A., Corbin, J., 1998. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Sage Publications, London, UK.
*Wall, T., et al., 2016. Developing evaluation indicators to improve the process of coproducing usable climate science. Weather. Clim. Soc. 9 (1), 95–107.
*Wellcome Trust, 2019. How We've Defined What Success Looks Like for Wellcome's Work. https://wellcome.ac.uk/news/how-weve-defined-what-success-looks-wellcomes-work.
*Wickson, F., Carew, A., 2014. Quality criteria and indicators for responsible research and innovation: learning from transdisciplinarity. J. Responsible Innov. 1 (3), 254–273.
Wyborn, C., et al., 2018. Understanding the impacts of research synthesis. Environ. Sci. Policy 86, 72–84.