



# Evaluating Research for Development: Innovation to Navigate Complexity

Marina Apgar<sup>1</sup> · Mieke Snijder<sup>1</sup> · Grace Lyn Higdon<sup>2</sup> · Sylvia Szabo<sup>3</sup>

Accepted: 6 February 2023 / Published online: 1 March 2023  
© The Author(s) 2023

## Abstract

Large publicly funded programmes of research continue to receive increased investment as interventions aiming to produce impact for the world's poorest and most marginalized populations. At this intersection of research and development, research is expected to contribute to complex processes of societal change. Embracing a co-produced view of impact as emerging along uncertain causal pathways often without predefined outcomes calls for innovation in the use of complexity-aware approaches to evaluation. The papers in this special issue present rich experiences of authors working across sectors and geographies, employing methodological innovation and navigating power as they reconcile tensions. They illustrate the challenges with (i) evaluating performance to meet accountability demands while fostering learning for adaptation; (ii) evaluating prospective theories of change while capturing emergent change; (iii) evaluating internal relational dimensions while measuring external development outcomes; (iv) evaluating across scales: from measuring local level end impact to understanding contributions to systems level change. Taken as a whole, the issue illustrates how the research for development evaluation field is maturing through the experiences of a growing and diverse group of researchers and evaluators as they shift from using narrow accountability instruments to appreciating emergent causal pathways within research for development.

**Keywords** Research for Development · Complexity · Innovation · Causal Pathways · Learning

## Résumé

Les grands programmes de recherche financés par des fonds publics continuent de recevoir des investissements accrus en tant qu'interventions visant à produire un im-

---

✉ Marina Apgar  
M.Apgar@ids.ac.uk

<sup>1</sup> Institute of Development Studies, Brighton, UK

<sup>2</sup> Monitoring, Evaluation and Learning Advisor, Independent, Eastbourne, UK

<sup>3</sup> Department of Social Welfare Counseling, Dongguk University, Seoul, South Korea



pour les populations les plus pauvres et les plus marginalisées dans le monde. À cette intersection entre la recherche et le développement, la recherche devrait contribuer aux processus complexes de changement sociétal. Pour adopter une vision coconstruite de l'impact comme phénomène émergent au fil de liens de causalité incertains, bien souvent sans résultats prédéfinis, il faut innover en utilisant des approches d'évaluation sensibles à la complexité. Les articles de ce numéro spécial présentent de riches expériences d'auteurs travaillant dans différents secteurs et zones géographiques, employant l'innovation méthodologique et le pouvoir de navigation tout en réconciliant les tensions. Ils illustrent les défis lorsqu'il s'agit (i) d'évaluer des performances pour répondre aux exigences de redevabilité tout en favorisant l'apprentissage pour l'adaptation; (ii) d'évaluer les théories prospectives du changement tout en saisissant le changement émergent; (iii) d'évaluer les dimensions relationnelles internes tout en mesurant les résultats de développement externes; (iv) d'évaluer à différentes échelles: de la mesure de l'impact final au niveau local à la compréhension des contributions au changement au niveau des systèmes. Pris dans son ensemble, ce numéro illustre la façon dont l'évaluation de la recherche pour le développement mûrit à travers les expériences d'un groupe toujours plus important et divers de chercheurs et d'évaluateurs qui abandonnent des outils de redevabilité étriqués afin d'apprécier les liens de causalité émergents au sein de la recherche pour le développement.

## The R4D Evaluation Challenge and Opportunity

The aim of research for development (R4D) is to use research as a vehicle to address critical development concerns, in order to improve the lives and livelihoods of disadvantaged communities across the world. R4D programmes are neither purely academic research nor are they discrete development interventions. As hybrid research endeavors they are expected to contribute to complex processes of societal change, and achievement of development outcomes in particular. Many large R4D programmes that are funded through international development (aid) budgets or philanthropic institutions are vast in size, scope, and ambition. Some have a long history, such as the CGIAR system of agricultural research which has received \$60 billion in investment over more than 40 years.<sup>1</sup> More recent forms of 'challenge driven' research include research aimed at the UN Sustainable Development Goals (SDGs) (Borrás 2019), the €80 billion<sup>2</sup> European Union Horizon 2020 research and development programme (Mazzucato et al. 2018), the €10 billion<sup>3</sup> WWForEurope project (Aiginger and Schratzenstaller 2016) and the £1.5 billion Global Challenges Research Fund (GCRF) funded by the United Kingdom government's Department of Business, Energy and Industrial Strategy (BEIS) (Barr et al. 2019) which includes 3000 awards within a highly diverse portfolio that spans across all SDGs and

<sup>1</sup> <https://www.cgiar.org/annual-report/performance-report-2020/assessing-cgiars-return-on-investment/>.

<sup>2</sup> [https://www.eeas.europa.eu/eeas/horizon-2020\\_en#:~:text=Horizon%202020%20is%20the%20EU's,\(roughly%20JPY%2010%20trillion\)](https://www.eeas.europa.eu/eeas/horizon-2020_en#:~:text=Horizon%202020%20is%20the%20EU's,(roughly%20JPY%2010%20trillion).).

<sup>3</sup> <https://cordis.europa.eu/project/id/290647>.



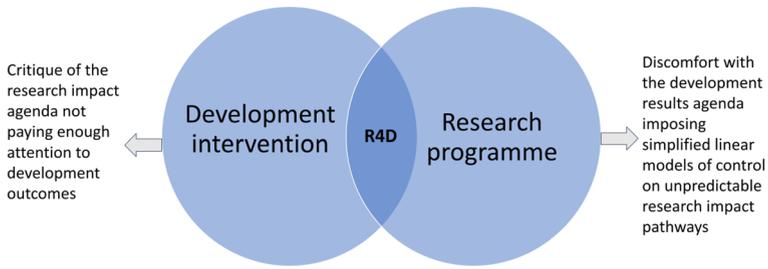


Fig. 1 R4D programmes sit at the intersection of research and development

lower-middle income countries (LMICs). The inter- and transdisciplinary research funded under these windows is framed around addressing ‘societal grand challenges’ requiring work across scales and sectors. At their core, R4D programmes are funded to undertake research as an intervention that produces direct, real-world impacts for the world’s poorest and most marginalized populations.

In today’s context of performance-based research funding, demand for evaluation of the impact of academic research has increased (Zacharewicz et al. 2019; Bornmann 2013) leading to ever more sophisticated metrics for assessing the excellence of research (Pinar and Horne 2022). The focus of these assessments, however, is premised on a linear pathway starting from new knowledge produced by excellent academic research, communicated through engagement activities and subsequently leading to changes in policy (see Georgalakis and Rose 2019 for broader debates on understanding how research leads to policy impact). Evaluation of agricultural research for development programmes have similarly been driven largely by assumptions of linear technology impact pathways that fail to engage with the complexity of outcomes that emerge through social interactions of multiple actors in agricultural systems (Belcher and Hughes 2021). This linear and overly simplified view of research impact pathways continues to dominate even as research is expected to engage with broader processes of change.

R4D funded via aid budgets experiences even higher levels of scrutiny on effectiveness given the prevalence of the development ‘results agenda’ (Eyben et al. 2015). As shown in Fig. 1 R4D programmes sit at the intersection of the research and development sectors creating pressures from both the research and the development impact agendas. A key task for evaluators is to reconcile research excellence with development effectiveness. On the one hand, part of the academic research community are uncomfortable with the imposition of a linear development evaluation orientation that doesn’t fully appreciate the unpredictability of research impact pathways<sup>4</sup> (Eyben et al. 2015). On the other hand, part of the development community, which is used to a results-based-management framing (e.g. Hatton and Schroeder 2007) finds the research excellence evaluation agenda not sufficiently focused on the harder to measure downstream

<sup>4</sup> See the LSE impact blog for a long-standing conversation on how to measure academic impact: <https://blogs.lse.ac.uk/impactofsocialsciences/the-handbook/chapter-6-is-there-an-impacts-gap-from-academic-work-to-external-impacts-and-perhaps-also-to-consequences-how-might-it-have-arisen-how-might-it-be-reduced/>.



development outcomes that move beyond policy change to real-world impact on people's lives and livelihoods (Peterson, <https://doi.org/10.1057/s41287-022-00565-7>).

Decades of R4D theory and practice has highlighted the complexity of the impact pathways of these hybrid programs (Horton and Mackay 2003; Thornton et al. 2017). They contain a multitude of actors that are engaged throughout the knowledge production process, moving away from a linear view of the discovery-to-application pipeline. Moving from simpler views of using the products of excellent research in knowledge exchange, to a co-produced view of impact along uncertain pathways, requires innovation in the way programmes are designed, operationalised, and consequently, evaluated (Blundo-Canto et al. 2017; Jacobi et al. 2020; Maru et al. 2018; Temple et al. 2018). Further, the intention around inclusion of marginalized voices and perspectives in how outcomes are achieved, requires systems approaches that open up opportunities for alternative pathways to emerge (Leach et al. 2007) with research sometimes acting as a disruptive force through which development is achieved (Ely et al. 2020).

Monitoring, Evaluation and Learning (MEL) systems that are fit for purpose for R4D programming must work with the complexity that arises from a large number of diverse partners working together in many and often integrated work streams, on problems in different contexts and on research that may have no predefined outcomes. This poses challenges to traditional evaluation designs that use a before and after (baseline-endline) logic, or seek counterfactual evidence of effectiveness, and requires acknowledgement of evaluation as embedded and enmeshed in complex social and political dynamics. The increased interest and funding for R4D has created an exciting opportunity to learn from experimentation with new evaluation designs and practices that are contributing to middle range theories (see Cartwright 2020 for a full explanation) of how R4D programmes work as well as evaluation theory and practice.

This special issue originated from a series of intentional learning exchanges by researchers and evaluators engaged in evaluation of programmes funded under the UKRI GCRF. The resulting special issue covers papers that shed light on different methods, approaches and areas of evaluation in R4D, including evaluation of specific funder portfolios, relational aspects of R4D, the link between R4D and SDGs and the use of theory of change and learning approaches. It is the first consolidated examination of R4D evaluation theory and practice, responding to the expansion of this field of funding and practice in the UK and beyond. In this introductory editorial we first discuss the R4D evaluation landscape as situated within a broader shift towards being more complexity-aware and better able to navigate uncertain impact pathways. We then describe four areas of tension that are experienced as challenges within complexity-aware evaluation practice and introduce the papers in this special issue by showing how they address each. Finally, we share our reflections on the future of R4D evaluation as editors of this special issue.

## The Shifting Landscape of Complexity-Aware Evaluation

The implications of complexity for examining if and how interventions, research among them, lead to societal changes (outcomes), are increasingly recognised in evaluation theory and practice across a number of fields (Walton 2016). In the



international development sector, embracing the SDGs as an overarching framework generated momentum around rethinking impact evaluation (e.g. Befani et al. 2014) to broaden beyond what until then had been narrow views of experimental designs as the ‘gold standard’. The so-called Stern review (Stern et al. 2012) commissioned by the UK Department for International Development (now the Foreign and Commonwealth Development Office) was pivotal in highlighting the need to nuance our understanding of methods through engaging with the underpinning frameworks used for making a causal claim. Subsequently, others have built on this foundation (see Gates and Dyson 2017; Jenal and Liesner 2017; Masset et al. 2021) to illustrate that evaluators working in conditions of complexity can choose from a range of approaches and methods, underpinned by distinct causal frameworks.<sup>5</sup> Table 1 illustrates the diversity in available designs.

In the social change sectors, including philanthropy, there is a noticeable turn to systems interventions or systems change strategies, leading to further theoretical and practice developments in the nascent field of evaluating systems change (Gates 2017; Hargreaves and Podems 2012; Lynn et al. 2021; Walton 2016). Related is the move towards innovation oriented and complexity informed programming (e.g. Burns and Worsley 2015; Jones 2011; Ramalingam 2013) calling for appropriate evaluation designs. We see growing demand and use of a family of evaluation approaches that are ‘complexity-aware’ including the well-known developmental evaluation (Patton 2010) and associated principles-focused evaluation (Patton 2017). They emphasize learning about how outcomes emerge along unpredictable impact pathways, with the intention of feeding learning back into implementation.

In the context of systemic, learning oriented and adaptive programming, there is no obvious evaluation design, and evaluators must work with programmers to choose and tailor appropriate designs from the variety of options available to them. Established guidance suggests methodological choice should be appropriate for the evaluation questions and the attributes of the intervention (HM Treasury; Befani 2020). And a number of typologies (e.g. Masset et al. 2021) and checklists (e.g. Bamberger et al. 2016) have been developed that acknowledge complexity should inform methodological choice. More recently, the term ‘bricolage’ is being used to guide evaluators in not only choosing and mixing methods, but to recombine different parts of methods (Aston and Apgar 2022; Hargreaves 2021) to support rigour in making causal claims amid complexity.

Another way in which evaluators are responding to complexity is the move towards greater use of theory-based approaches which start with an articulated theory of how an intervention is thought to achieve impact (Rogers and Weiss 2007; Weiss 1997). A wide range of methods (see Table 1) fit within this family and have particular ways of developing theory, causal assumptions and testing or refining them through evaluation research. A core contribution of these approaches, and in particular realist evaluation (Pawson and Tilley 1997) and process tracing (Beach 2017; Stachowiak et al. 2020) is acknowledging contextual conditions as part of the

<sup>5</sup> See Stern et al. (2012) for a full exploration of the most common causal frameworks (counterfactual, regularity, configurational and generative).



**Table 1** Evaluation approaches and methods and their causal frameworks

Approach	Examples of methods	Underpinning causal framework
Theory-based approaches	<ul style="list-style-type: none"> <li>• Contribution analysis</li> <li>• Process tracing</li> <li>• Realist evaluation</li> <li>• General elimination methodology</li> <li>• Qualitative impact assessment protocol</li> <li>• Innovation history</li> </ul>	In-depth theoretical analysis of causal processes or mechanisms in context. Many of these methods are grounded in a generative causal framework, though configurational frameworks are also relevant to some
Participatory approaches	<ul style="list-style-type: none"> <li>• Most significant change</li> <li>• Outcome harvesting</li> <li>• Collaborative outcomes reporting</li> <li>• Collaborative yarning</li> </ul>	Engagement of change agents (programme participants) in defining the scope of change to be explored and in analysis of causal claims. Many of these methods are grounded in a generative causal framework
Case-based approaches	<ul style="list-style-type: none"> <li>• Within-case</li> <li>• Across-case</li> </ul>	Analysis of causal processes within a case or across multiple cases based largely on a configurational causal framework though generative frameworks are also relevant (especially when combined with theory-based approaches)

Adapted from Jenal and Liesner (2017), Lynn et al. (2021, Lynn and Apgar (forthcoming))

causal relationships under investigation. Underpinned by configurational and generative causal frameworks, they are well suited to answer evaluation questions about not just what, but how, in what conditions and for whom are outcomes and impact achieved. Further, these approaches can support learning through iterative use of theory of change (ToC), as illustrated by evolving approaches to contribution analysis (Apgar et al. 2020; Ton et al. 2019).

A common theme across these learning-oriented approaches to evaluating large and complex programmes is to appreciate evaluation design as an iterative process rather than a single decision point at the outset. Evaluators working in conditions of complexity must evolve their designs as outcomes emerge and assumptions about causal links are clarified, and as learning agendas are reshaped along the way through collaboration with stakeholders. This has implications for the capacities required as evaluators must shift their role from being external technical experts to embedded facilitators (Barnett and Eager 2021). The move away from evaluation as simply a technical endeavor, to embracing the politics within the evaluation process and how it informs decision making (Eyben et al. 2015; Polonenko 2018) is reopening long standing debates about whose knowledge counts in evaluation (see Estrella et al. 2000). Related calls for greater equity-orientation in evaluation (Forrestieri 2020; Gates et al. 2022; Hall 2020) are focusing attention on power ‘in’ and ‘of’ evaluation (Hanberger 2022). These trends are creating new opportunities for re-centering the role of the evaluator as a knowledge broker engaging with commissioners, programmers and change actors within systems.

Evaluating R4D programmes sits at the intersection of research and development, and takes place within an established yet still evolving landscape of learning oriented and complexity-aware evaluation practice. Exploring the uncertain impact pathways of R4D requires navigating difficult methodological choices and managing across distinct, at times contested fields of practice and thinking enmeshed in politics and power. The papers in this issue help us to understand what these tensions look like through experiences of diverse teams working at different scales and across contexts. In the following section we introduce the papers through four interconnected tensions, grounding each in the literature and highlighting how papers in this issue contribute to these areas of contestation, debate and praxis within the field of R4D evaluation.

## **Navigating Tensions in R4D Evaluation**

We organize our introduction to the papers in this special issue by describing four areas of tension that are experienced as challenges within complexity-aware evaluation practice. The four areas span the realms of decision making span the realms of decision making on focus and purpose and related decisions on appropriate methodological choices, all the way through to the way in which evaluation findings are used and the strength of causal claims are assessed. While the four areas of tensions are in practice interconnected, for ease of presentation we explore each separately



and introduce the papers featured in this issue as they contribute to R4D evaluation theory and practice.

### **Evaluating Performance to Meet Accountability Demands While Fostering Learning for Adaptation**

The tension between learning and accountability is well recognised in the context of development evaluation (Estrella et al. 2000; Guijt 2010; Guijt and Roche 2014) and links to broader debates around the politics of evidence and the narrow framing of the ‘results agenda’ in international development (Eyben et al. 2015). This tension is particularly relevant in how the effectiveness of newly developed hybrid challenge-driven research programmes, which are implemented by academic institutions and funded through aid budgets, is understood and consequently, how programmes are evaluated. A strict performance management and narrow accountability focus—be that on research excellence or development outcomes—can inhibit open and honest sharing of successes and failures and shut down the space required for learning to drive the generation of innovation—one of the aims of R4D. These poor conditions are brought into relief through technical challenges, such as traditionally used methods for performance monitoring and management, which rely on linear and predefined models of performance, which focus solely on academic output, combined with political challenges inherent within different stakeholders’ epistemological beliefs, particular agendas, and influence within a given program that shape or winnow the type of learning that can be generated (Aston et al. 2021).

For years—and to varying degrees of success—communities of practice seeking to understand complex, learning-oriented initiatives have included stakeholders interested in measuring policy and advocacy, thinking and working politically, and adaptive management. Practitioners have been debating approaches and practicing methods to navigate these tensions. Activities embedded within complexity-aware evaluation designs (see Table 1) include structured reflective moments conducted in intervals that feed into decisions made during implementation, thus providing documentation and justification for ‘real-time’ adjustments which address accountability requirements while simultaneously providing opportunities for program stakeholders to take stock and collectively learn. Other activities include capturing observed progress toward outcome level changes during implementation through participatory data collection and synthesis processes which provide evidence to understand program performance while also allowing program stakeholders to appreciate what and how outcomes are emerging (Laws and Marquette 2018; Pasanen and Barnett n.d.; Reisman et al. 2007).

From decades of experience we know that the space for learning requires more than simply having the right tools. The political conditions must also be favorable: participatory and qualitative evidence have to be considered valid, reporting failure and acknowledgement of uncertain predictions must be considered acceptable, and course corrections-accompanied by diverting funds-must be allowed. In recent years, while uptake of more learning-oriented approaches has grown, the efficacy of those approaches can be drastically limited by the program’s funding environment if



the environment remains premised on linear programme modalities with predicted milestones (McCullough et al. 2017).

Few tools illustrate this juxtaposition better than the recent uptake of the use of ToC over the past decade (Vogel 2012). For some, the use of ToC represents a mechanism of resistance to strict indicator-based results frameworks, allowing for more collaborative, nuanced and non-linear predictive program modeling that does not pin programs down to milestones (Apgar et al. 2022). In others' experience, ToC has been co-opted and is yet another version of a results matrix (e.g. logframe 2.0), especially in cases where funders require performance predictions to be tied to the ToC at the outset of the program and have low appetite for adaptation of the ToC during implementation.

Chapman et al. (this issue) explore the role of ToC in navigating the tensions between learning and accountability in R4D evaluation through examination of the use of visual ToCs by MEL practitioners and evaluators working on ten large R4D programmes funded under the GCRF. The authors conclude across all experiences, that due to the performance management requirements- including output-focused results reporting required by the funder—most of the 10 programmes opted to simplify their visualizations and overall use of ToCs. Although some set out to use ToCs as strategic learning tools, due to capacity issues, resource constraints, and the political operating environment of these R4D programmes, simplified ToCs were found to be most useful to establish standardized language across multi-disciplinary teams, depict 'big picture' programmatic goals, and inform logical framework design—not as learning tools. The paper highlights how the accountability environment of R4D programmes may challenge the intention to use ToC as a learning tool.

Apgar et al. (<https://doi.org/10.1057/s41287-023-00576-y>) explore the use of Social Network Analysis (SNA) in three large scale R4D programmes as a tool for learning about how the programme structures evolved through time (linked to performance and programme assumptions in ToC) while also using the learning to intentionally weave the networks in desirable directions (adaptation). Through the three case experiences they reveal tensions between these two purposes especially in how SNA findings are interpreted and by whom, and conclude that “navigating the challenges of interpretation and ethical dilemmas requires careful consideration as well as an enabling institutional and political environment for use of SNA to support learning.” (Apgar et al. <https://doi.org/10.1057/s41287-023-00576-y> pxy). They recommend embedding the interpretation of SNA findings into participatory sense making moments within the broader adaptive management designs of R4D programmes so as to hold open sufficient space for learning to be actioned.

Embracing complexity and moving away from standardized and predefined accountability metrics requires R4D evaluation stakeholders to decide first what it is they value, in order to build an appropriate approach to learn as outcomes emerge. Peterson (<https://doi.org/10.1057/s41287-022-00565-7>) shares an alternative to 'Value for Money', one of the main instruments of accountability, particularly prevalent in the current constrained UK government spending environment. Peterson identifies the need to bridge between standard econometric evaluation designs and complexity-aware methodologies to build an approach that is fit for purpose. She



presents a collaboratively developed rubric-based approach to reviewing R4D projects, couched in a constructivist paradigm that allows for valuing both the process and the outcomes of research, and reflects on its application in two large R4D portfolios. The paper illustrates how accountability and learning can be brought together through methodological innovation.

### **Evaluating Prospective Theories of Change While Capturing and Exploring Emergent Change**

Building on the use of ToC as a learning and management tool, is the now common practice of using ToC to inform theory-based approaches to impact evaluation (see Table 1). Yet being theory based does not necessarily lead to being complexity-aware. Indeed, there is often a tension experienced between prospective and retrospective use of ToC, with a separation of monitoring tools to track predefined indicators within a ToC that looks forward and evaluation aiming to understand results looking backwards. As Jenal and Liesner (2017) note, a key criticism of prospective approaches to ToC in evaluating systems change is that they do not capture unexpected changes. Doing deep causal thinking at the outset of a programme and detailing causal theories of change is helpful to guide evaluation research and zoom into specific causal links within a ToC. But, R4D impact pathways are complex and unpredictable, leading often to unexpected effects which prospective use of ToC may miss if applied in an overly linear way. In particular, the visualization of ToC bears the risk of oversimplification and the unintended effect of pushing a linear causal logic that crowds out any space for emergence (Davies 2018; Wilkinson et al. 2021).

Chapman et al. (this issue) reflect on how to balance funder requirements for a simplified ToC with one that embraces the complexity of R4D and leaves space for emergence. They share two strategies to navigate this tension. Firstly, some of the programmes used nested ToCs, with a linear overarching ToC that was then broken down into specific ToCs detailing how a specific component of the R4D programme was contributing to specific outcomes and impacts or how outcomes and impacts were going to be achieved in a specific geographical area of the programme. These nested ToC were more adaptable and manageable than the large programme level ToC and gave the programme space to embrace more uncertainty and emergence in the evaluation designs. Secondly, one programme used visualized impact pathways combining system mapping with ToC and participatory approaches that allowed for iterative revisions, non-linearity and feedback loops. While a resource intensive approach, it illustrates the need to innovate with ToC when working with large complex projects to keep the space open for the exploratory nature of R4D.

An additional challenge with prospective approaches to ToC is that they can suffer from overconfidence in projecting a single anticipated future that hides other contributory factors. This is a form of confirmation bias often critiqued when evaluating complex processes of change. The Impact Weaving method introduced by Blundo-Canto et al. (<https://doi.org/10.1057/s41287-022-00566-6>) is a way to embrace complexity in ToC by updating researcher-generated impact pathways of



agricultural innovations with other stakeholders' knowledge. This method plays with prospective and retrospective aspects of the impact pathways under investigation by combining past knowledge from researchers and other stakeholders, with knowledge of the current situation and context from stakeholders who are or will be using the innovations, with future knowledge through visioning and participatory scenario building. This transdisciplinary method creates contextually relevant impact pathways that provide a more systemic, triangulated and grounded vision of how change actually happens.

Often combined with prospective approaches to ToC in theory-based evaluation are goal independent evaluation methods such as Outcome Harvesting (Wilson-Grau 2018, p.) and Most Significant Change (Davies and Dart 2005) that capture emergent change by identifying outcomes after they have been produced, and tracing contribution back to the programme under evaluation. These are forms of causes-of-effects analysis often found in case-based methods (Goertz and Mahoney 2012). In this issue, a new method that aims to combine prospective and retrospective approaches in the context of policy evaluations is introduced by Douthwaite et al. (<https://doi.org/10.1057/s41287-022-00569-3>) called outcome trajectory evaluation (OTE). The aim of OTE is to cover all factors that are hypothesized to be influenced by policy outcomes, not just the ones that are targeted by the policy. In OTE outcomes are “understood to emerge in complex adaptive systems, through the interaction of actors, their strategies and decision-making, institutions, artifacts (i.e., technology) and knowledge” (Douthwaite et al. <https://doi.org/10.1057/s41287-022-00569-3>). This view of outcomes embraces emergence and systems change at its core by taking a long-term view of outcomes, rather than seeing outcomes as a single episode of change. This offers a more balanced view of the contribution of any specific intervention within context.

### **Evaluating Internal Relational Dimensions While Measuring External Development Outcomes**

R4D programmes, through their hybrid nature, require collaboration across diverse actors and sectors—from research institutions and researchers of different disciplines to development agencies, government departments and practitioners—to focus together on addressing grand societal challenges (or SDGs) often working across contexts. These large inter- or transdisciplinary collaborations create relational spaces through which impact is enabled downstream, often taking a long time to materialize. This has been referred to as the ‘productive interactions’ space between research and other societal actors and has been informing the evaluation of European research systems (Muhonen et al. 2020).

In these large R4D collaborations, it is such productive interactions that create the conditions for outcomes to emerge, rather than achievements of individual actors leading to discrete outcomes (Hargreaves 2021; Walton 2016). The relational components are inherently unpredictable, requiring, as noted already, an emphasis on learning real time as relationships evolve and opportunities for impact become clearer. Building on the tension already discussed around accountability versus



learning, a focus solely on measuring ‘results’ downstream misses the opportunity to understand the conditions through which they emerge. Within complexity aware evaluation approaches, evaluation questions of how and why outcomes are emerging are well placed to inquire into the relational as mechanisms for achieving impact.

Specific approaches to exploring these relational dynamics are gaining ground, for example, appreciating the processes of co-production in interdisciplinary research programmes (de Sandes-Guimarães et al. 2022). Processes of co-production are premised on productive and equitable collaborations across actors—in other words the internal ‘ways of working’. Snijder et al. (<https://doi.org/10.1057/s41287-023-00578-w>) focus specifically on evaluation and learning of equitable partnerships through a comparison across five large scale R4D programmes. Across the five cases they illustrate how decolonial, feminist and participatory approaches were used to address hard to shift power asymmetries related to: funding flows from the so-called ‘global north’ to lower and middle income countries; hierarchies between senior researchers and early career researchers as well as across disciplines and genders. The authors argue that participatory approaches embedded in programme MEL allow internal power dynamics to be revealed and acted upon in support of adaptive management, and they propose a framework that distills key principles for evaluating equitable partnerships in R4D programmes.

Looking across the whole GCRF in evaluating the early phases of implementation, Vogel and Barnett (<https://doi.org/10.1057/s41287-023-00579-9>) share emerging evidence on how conditions for R4D impact are built, shining a light on the processes of set up and implementation as mechanisms for impact down the road. They identify four building blocks as “elements and processes that projects need to build into their research to position it for impact.” (Vogel and Barnett, <https://doi.org/10.1057/s41287-023-00579-9>, p. X): (1) Scoping of development issues with stakeholders on the ground for relevance; (2) Fair and equitable partnerships between partners in the so-called Global North and lower and middle income countries, including non-academic partners, integrating mutual capacity building; (3) Gender, social inclusion and poverty prioritized in policies and implementation, and (4) Stakeholder engagement in lower and middle income countries to support positioning research for use. In using the GCRF as a case of R4D they echo Snijder et al. (<https://doi.org/10.1057/s41287-023-00578-w>) by highlighting the efforts and success of the signature GCRF Hubs in setting up mechanisms to ensure fair representation of partners within internal governance and decision-making as well as evaluation of equity in these partnerships.

This focus on the internal dynamics of large multidisciplinary research programmes is leading to use of novel methods in evaluating research teams and how they work, such as SNA (e.g. (Higgins and Smith 2022). Apgar et al. (<https://doi.org/10.1057/s41287-023-00576-y>) zoom into the relational structures of R4D programmes through evaluating them as network building initiatives. The SNA method, employed within MEL systems designed to inform adaptive management, revealed how the network structures matched the expectations of a centralized setup of the programmes with central coordination teams in the Global North. It also brought to light surprising evolutions in the structures that enabled questioning of underlying assumptions, for example, around the gendered dynamics of collaboration. The



authors suggest that paying more attention to the relational at the outset, through developing ‘contextualized theories of collaboration’ would enhance the use of SNA in evaluation through guiding studies in more purposeful ways.

### **Evaluating Across Scales: From Measuring Local Level End Impact to Understanding Contributions to Systems Level Change**

Another lens through which tensions can be felt and analyzed in R4D programmes relates to the scales at which impact is theorized to emerge and consequently evaluated. Like complex systems, R4D programmes include multiple levels of ‘nested’ interventions within them. To inform future investments at the fund level, portfolios of projects are evaluated to assess what has been achieved overall and how large scale interventions work, and within a fund, individual programmes or projects of varying sizes focused on particular challenges in contextualized ways and are evaluated to assess their outcomes and impact and to learn how to improve R4D interventions. R4D funds are increasingly framed around the SDGs suggesting ‘global level’ impact as the end goal at the portfolio level. Yet these global goals need to be materialized in concrete changes in people’s lives and livelihoods—such as poverty reduction—which point to measuring real change in specific locations linked to individual interventions. The levels are interconnected, creating unique challenges for evaluation design.

In the context of the UK, the Newton Fund and the GCRF evaluations offered opportunity to innovate, connecting across project and fund level evaluations. Two papers in this issue share learning from these evaluations as informative cases of large R4D portfolios. Peterson (<https://doi.org/10.1057/s41287-022-00565-7>) details the steps used to build a methodology for portfolio level assessment of Value for Money based on multiple-case studies of individual projects. The cases provide in-depth details of R4D impact, which enables comparison across cases while being attentive to the uniqueness and contextuality of each case. Rubrics were developed with all stakeholders to agree what criteria to value, a method that allows tailoring to the specific needs of R4D programmes with criteria such as ‘equitable partnerships’, and ‘likelihood of fund level impact’ to allow connecting across the project and fund levels. Vogel and Barnett (<https://doi.org/10.1057/s41287-023-00579-9>) share emerging evidence from the GCRF fund level evaluation which is still ongoing. The evaluation team in the early phases had to navigate the diversity produced by a highly devolved structure of the fund which is delivered through 17 partners through existing research and University systems. The resulting 3000 awards make up a highly diverse portfolio that spans across all SDGs and LMICs. The ‘building blocks’ for impact they identify are the result of synthesis across many case studies through their modular evaluation design and provide insights on core mechanisms for impact of R4D programmes.

The evaluation community has been grappling with the SDG agenda and pushing towards evaluation being at the service of transformative change in systems (Aronsson and Hassnain 2019)). The challenge of scale is evident in the



complexity of setting SDG targets and indicators. Some experts have suggested a two-track solution to measuring SDGs would be most efficient, allowing for both ‘goal level’ measures and lower level ‘technical indicators’ that could relate to interventions more directly (Davis et al. 2015). Gonzales et al. (<https://doi.org/10.1057/s41287-022-00573-7>) engage with the challenge of measuring contribution of an R4D project on trade and environment. They propose a methodology for mapping specific contributions of the project through its planned outcomes areas (named ‘big wins’) to the SDG targets. This detailed method links to the evaluation design, and is able to provide a clear line of sight between outputs of every work package within the project to system level measures of impact.

## What Next for R4D Evaluation?

We have described R4D programmes as sitting at the intersection of research and development (see Fig. 1) and the papers in this issue present rich experiences of a community of evaluation practitioners and researchers that are innovating methodologically to navigate and reconcile the tensions that arise from this hybrid reality. As Vogel and Barnett (<https://doi.org/10.1057/s41287-023-00579-9>) note, the GCRF evaluation is framing this hybrid reality through a unifying construct of ‘development excellence’ as the main goal of R4D programmes. We understand R4D evaluation as an evolving field and this issue is the result of evaluators, programme managers and researchers exploring the practical implications of navigating these tensions, often with little formal guidance and facing many hurdles along the way. This growing community of R4D evaluation researchers is represented in this special issue by over 40 co-authors from over 10 countries in Europe and middle- and low-income countries, including a significant number of early career researchers. Through the papers, the voices of this community come to life, acknowledging the embeddedness of R4D evaluation, and the reflexivity required of diverse evaluation teams.

Taken together, the papers in this issue provide insights on how complexity-aware and ‘bricolaged’ evaluation designs are implemented in the context of R4D programmes. In the process of developing this special issue, we, the editorial team, and all co-authors engaged in the GCRF programmes experienced a major funding crisis that served to sharpen our empirical understanding of the challenge of engaging in the contested spaces we have described here as areas of tension. In November 2020, in response to the COVID-19 pandemic, the UK Aid budget was significantly reduced, and as a consequence the funding of the 12 signature GCRF Hubs was cut by up to 70% (UKRI 2021; Nwako et al. 2023). For a period, the implementation teams faced high levels of uncertainty about future funding and the viability of the evaluation research. This led to a sudden shift away from the learning orientation that had informed evaluation designs and practice. Additional reporting requirements were placed upon all programmes, with new hoops to be jumped through that forced a reorientation on proving what had been achieved and away from exploring the ways in which impact opportunities were emerging along unpredictable pathways. The space for deepening learning around



the relational and internal dynamics was curtailed and the reduced budgets meant in some cases MEL systems and expertise was no longer a priority.

This recent GCRF experience highlights the double-edged sword of evaluation as a performance management tool and a research and learning tool. And it is not entirely unique. Indeed, evaluators working within the CGIAR system have at times experienced similar shifts in the use of evaluation in response to reduced funding. Douthwaite et al. (2017) describe one such experience which led to a shift away from systems-oriented research programmes, arguing that underpinning the shift was a different way of valuing R4D programmes, based on a narrow view of causal claims requiring counterfactual designs. As Peterson et al. (<https://doi.org/10.1057/s41287-022-00565-7>) show, there is still room to deepen the shift away from the underlying positivist leaning in the accountability instruments used to assess the value of research to see greater uptake of methods that are more fit for purpose.

Despite, or perhaps because of, the challenges that remain, the papers in this issue illustrate that reconciling tensions in R4D evaluation is possible, offering methodological innovations that show in practice that the trend towards broadening evaluation designs to embrace complexity is gaining momentum. Related calls in the sustainability field for greater funding flexibility to stay the course, to give time and space for the impacts of systemic and transdisciplinary research to materialize downstream (e.g. Ely 2021; Benedum et al. 2022) and testing of funder developed approaches to evaluating research such as the IDRC Research Quality+ tool (McLean et al. 2022; Lebel and McLean 2018) provides further grounds for optimism. The field of R4D evaluation will continue to mature and this issue illustrates that diverse experiences and learning across R4D stakeholders, including researchers, research managers, evaluators and funders is contributing to its coming of age.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Aiginger, K., and M. Schratzenstaller. 2016. New dynamics for Europe: Reaping the benefits of socio-ecological transition. Synthesis report part I (research report no. 11). WWForEurope Deliverable. <https://www.econstor.eu/handle/10419/169308>.
- Apgar, M., K. Hernandez, and G. Ton. 2020. Contribution analysis for adaptive management, Briefing Note, London: Overseas Development Institute, 14.
- Apgar, M., M. Snijder, P. Prieto Martin, G. Ton, S. Macleod, S. Kakri, and S. Paul. 2022. Designing contribution analysis of participatory programming to tackle the worst forms of child labour.



- CLARISSA research and evidence paper 2, brighton: institute of development studies. <https://doi.org/10.19088/CLARISSA.2022.003>
- Aronsson, I.-L., and H. Hassnain. 2019. Value-based evaluations for transformative change. In *Evaluation for Transformational Change: Opportunities and challenges for the Sustainable Development Goals*, (eds.) Rob D. van den Berg, Cristina Magro and Silvia Salinas Mulder. Exeter, UK: IDEAS. 89.
- Aston, T., and M. Apgar. 2022. *The Art and Craft of Bricolage in Evaluation*, CDI Practice Paper 24, Brighton: Institute of Development Studies. <https://doi.org/10.19088/IDS.2022.068>.
- Aston, T., C. Roche, M. Schaaf, and S. Cant. 2021. Monitoring and evaluation for thinking and working politically. *Evaluation* 28.1: 36–57. <https://doi.org/10.1177/13563890211053028>.
- Bamberger, M., J. Vaessen, and E. Raimondo. 2016. Complexity in development evaluation. Dealing with complexity in development evaluation: A practical approach. Thousand Oaks, CA: SAGE Publications, 1–25.
- Barnett, C., and R. Eager. 2021. Evidencing the impact of complex interventions: The ethics of achieving transformational change. In *Ethics for Evaluation*, 124–140. London: Routledge.
- Barr, J., P. Simmonds, B. Bryan, and I. Vogel. 2019. *Inception Report—Global Challenge Research Fund (GCRF) Evaluation—Foundation Stage*. <https://doi.org/10.13140/RG.2.2.31983.89762>.
- Beach, D. 2017. Process-tracing methods in social science. In Thompson WR (ed) *Oxford Research Encyclopedia of Politics: qualitative political methodology*, Oxford University Press, Oxford. <https://doi.org/10.1093/acrefore/9780190228637.013.176>
- Befani, B. 2020. Choosing appropriate evaluation methods—a tool for assessment and selection (version two), guildford: centre for the evaluation of complexity across the nexus
- Befani, B., C. Barnett, and E. Stern. 2014. Introduction—Rethinking impact evaluation for development. *IDS Bulletin* 45 (6): 1–5. <https://doi.org/10.1111/1759-5436.12108>.
- Belcher, B.M., and K. Hughes. 2021. Understanding and evaluating the impact of integrated problem-oriented research programmes: Concepts and considerations. *Research Evaluation* 30 (2): 154–168. <https://doi.org/10.1093/reseval/rvaa024>.
- Benedum, M., B.E. Goldstein, A. Ely, M. Apgar, L. Pereira, and D. Manuel-Navarrete. 2022. Lessons for transformations organizations from the pathways network: A transformations community dialogue. *Social Innovations Journal* 15. <https://socialinnovationsjournal.com/index.php/sij/article/view/4974>.
- Blundo-Canto, G., P. Läderach, J. Waldock, and K. Camacho. 2017. Learning through monitoring, evaluation and adaptations of the “ *Outcome Harvesting* ” tool. *Cahiers Agricultures* 26 (6): 65004. <https://doi.org/10.1051/cagri/2017054>.
- Borrmann, L. 2013. What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology* 64 (2): 217–233.
- Borrás, S. 2019. Domestic capacity to deliver innovative solutions for grand social challenges. In *The Oxford Handbook of Global Policy and Transnational Administration*. <https://doi.org/10.1093/oxfordhb/9780198758648.013.42>.
- Burns, D., and S. Worsley. 2015. *Navigating complexity in international development: Facilitating sustainable change at scale*. Warwickshire: Practical Action Publishing.
- Cartwright, N. 2020. Middle-range theory. *Theoria: An International Journal for Theory, History and Foundations of Science* 35 (3): 269–323.
- Davies, R. 2018. Representing theories of change: Technical challenges with evaluation consequences. *Journal of Development Effectiveness* 10 (4): 438–461.
- Davies, R., and J. Dart. 2005. The ‘most significant change’ (MSC) technique. In *A Guide to Its Use*. <http://www.mande.co.uk/docs/MSCGuide.htm>. Accessed 25 February 2023
- Davis, A., Z. Matthews, S. Szabo, and H. Fogstad. 2015. Measuring the SDGs: A two-track solution. *The Lancet* 386 (9990): 221–222.
- de Sandes-Guimarães, L.V., R. Velho, and G.A. Plonski. 2022. Interdisciplinary research and policy impacts: Assessing the significance of knowledge coproduction. *Research Evaluation* 31 (3): 344–354. <https://doi.org/10.1093/reseval/rvac008>.
- Douthwaite, B., J.M. Apgar, A.-M. Schwarz, S. Attwood, S. Senaratna Sellamuttu, and T. Clayton. 2017. A new professionalism for agricultural research for development. *International Journal of Agricultural Sustainability* 15 (3): 238–252.
- Ely, A. 2021. *Transformative Pathways to Sustainability: Learning Across Disciplines, Cultures and Contexts*. London: Routledge.



- Ely, A., A. Marin, L. Charli-Joseph, D. Abrol, M. Apgar, J. Atela, R. Ayre, R. Byrne, B.K. Choudhary, V. Chengo, A. Cremaschi, R. Davis, P. Desai, H. Eakin, P. Kushwaha, F. Marshall, K. Mbeva, N. Ndege, C. Ochieng, ..., L. Yang. 2020. *Structured Collaboration Across a Transformative Knowledge Network—Learning Across Disciplines, Cultures and Contexts? Sustainability*, 12 (6): 2499.
- Estrella, M., J. Blauert, D. Campilan, J. Gaventa, J. Gonsalves, I.M. Guijt, D.A. Johnson, and R. Ricafort. 2000. *Learning From Change: Issues and Experiences in Participatory Monitoring and Evaluation*. IDRC.
- Eyben, R., I. Guijt, C. Roche, and C. Shutt, eds. 2015. *The Politics of Evidence and Results in International Development: Playing the Game to Change the Rules?* Warwickshire: Practical Action Publishing Ltd. <https://doi.org/10.3362/9781780448855>.
- Forestieri, M. 2020. Equity implications in evaluating development aid: The Italian case. *Journal of MultiDisciplinary Evaluation* 16 (34): 65–90.
- Gates, E.F. 2017. Learning from seasoned evaluators: Implications of systems approaches for evaluation practice. *Evaluation* 23 (2): 152–171. <https://doi.org/10.1177/1356389017697613>.
- Gates, E., and L. Dyson. 2017. Implications of the changing conversation about causality for evaluators. *American Journal of Evaluation* 38 (1): 29–46.
- Gates, E.F., J. Madres, J.N. Hall, and K.B. Alvarez. 2022. It takes an ecosystem: Socioecological factors influencing equity-oriented evaluation in New England, U.S., 2021. *Evaluation and Program Planning* 92: 102068. <https://doi.org/10.1016/j.evalprogplan.2022.102068>.
- Georgalakis, J., and P. Rose. 2019. Introduction: identifying the qualities of research-policy partnerships in international development—A new analytical framework. *IDS Bulletin*. <https://doi.org/10.19088/1968-2019.103>.
- Goertz, G., and J. Mahoney. 2012. *A Tale of Two Cultures*. Princeton: Princeton University Press.
- Guijt, I.M. 2010. Accountability and learning: Exploding the myth of incompatibility between accountability and learning. In *NGO Management*, 339–352. <https://doi.org/10.1201/9781849775427-36>.
- Guijt, I., and C. Roche. 2014. Does impact evaluation in development matter? Well, it depends what it's for! *The European Journal of Development Research* 26 (1): 46–54. <https://doi.org/10.1057/ejdr.2013.40>.
- Hall, M.E. 2020. Blest be the tie that binds. *New Directions for Evaluation* 2020 (166): 13–22. <https://doi.org/10.1002/ev.20414>.
- Hanberger, A. 2022. Power in and of evaluation: A framework of analysis. *Evaluation* 28 (3): 265–283.
- Hargreaves, M. 2021. Bricolage: A pluralistic approach to evaluating human ecosystem initiatives. *New Directions for Evaluation* 2021 (170): 113–124. <https://doi.org/10.1002/ev.20460>.
- Hargreaves, M.B., and D. Podems. 2012. Advancing systems thinking in evaluation: A review of four publications. *American Journal of Evaluation* 33 (3): 462–470. <https://doi.org/10.1177/1098214011435409>.
- Hatton, M.J., and K. Schroeder. 2007. Results-based management: Friend or foe? *Development in Practice* 17 (3): 426–432.
- Higgins, L.E., and J.M. Smith. 2022. Documenting development of interdisciplinary collaboration among researchers by visualizing connections. *Research Evaluation* 31 (1): 159–172. <https://doi.org/10.1093/reseval/rvab039>.
- Horton, D., and R. Mackay. 2003. Using evaluation to enhance institutional learning and change: Recent experiences with agricultural research and development. *Agricultural Systems* 78 (2): 127–142. [https://doi.org/10.1016/S0308-521X\(03\)00123-9](https://doi.org/10.1016/S0308-521X(03)00123-9).
- Jacobi, J., A. Llanque, S. Bieri, E. Birachi, R. Cochard, N.D. Chauvin, C. Diebold, R. Eschen, E. Frossard, T. Guillaume, S. Jaquet, F. Kämpfen, M. Kenis, D.I. Kiba, H. Komarudin, J. Madrazo, G. Manoli, S.M. Mukhovi, V.T.H. Nguyen, C. Pomalègni, S. Rüegger, F. Schneider, N. Tri-Dung, P. von Groote, M.S. Winkler, J.G. Zaehring, and C. Robledo-Abad. 2020. Utilization of research knowledge in sustainable development pathways: Insights from a transdisciplinary research-for-development programme. *Environmental Science & Policy* 103: 21–29. <https://doi.org/10.1016/j.envsci.2019.10.003>.
- Jenal, M., and M. Liesner. 2017. Causality and attribution in market systems development: Report.
- Jones, H. 2011. *Taking Responsibility for Complexity: How Implementation Can Achieve Results in the Face of Complex Problems*. London: Overseas Development Institute.
- Laws, E., and H. Marquette. 2018. Thinking and working politically: Reviewing the evidence on the integration of politics into development practice over the past decade. *Thinking and Working Politically Community of Practice*.



- Leach, M., I. Scoones, and A. Sterling. 2007. *Pathways to Sustainability: An Overview of the STEPS Centre Approach. STEPS Approach Paper*. STEPS Centre.
- Lebel, J., and R. McLean. 2018. A better measure of research from the global south. *Nature*. <https://doi.org/10.1038/d41586-018-05581-4>.
- Lynn, J., S. Stachowiak, and J. Coffman. 2021. Lost causal: Debunking myths about causal analysis in philanthropy. *The Foundation Review*. <https://doi.org/10.9707/1944-5660.1576>.
- Lynn, J and M. Apgar. (forthcoming). Designs and analytical methods for exploring causality amid complexity. Chapter in Newcomer and Mumford. Editors. *Research Handbook on Programme Evaluation*. Edwar Elgar Publishing.
- Maru, Y.T., A. Sparrow, J.R.A. Butler, O. Banerjee, R. Ison, A. Hall, and P. Carberry. 2018. Towards appropriate mainstreaming of “Theory of Change” approaches into agricultural research for development: Challenges and opportunities. *Agricultural Systems* 165: 344–353. <https://doi.org/10.1016/j.agsy.2018.04.010>.
- Masset, E., S. Shrestha, and M. Juden. 2021. *Evaluating Complex Interventions in International Development*. *CEDIL Methods Working Paper 6*. London: Centre of Excellence for Development Impact and Learning (CEDIL). <https://doi.org/10.51744/CMWP6>.
- Mazzucato, M., European Commission, and Directorate-General for Research and Innovation. 2018. *Mission-Oriented Research & Innovation in the European Union: A Problem-Solving Approach to Fuel Innovation-Led Growth*. Publications Office of the European Union.
- McCulloch, N., & Piron, L.-H. (2019). Thinking and working politically: Learning from practice. Overview to Special Issue. *Development Policy Review*, 37 (S1), O1–O15. <https://doi.org/10.1111/dpr.12439>.
- McLean R., Z. Ofir, A. Etherington, M. Acevedo, and O. Feinstein. 2022. *Research Quality Plus (RQ+)—Evaluating Research Differently*. Ottawa: International Development Research Centre (IDRC). <https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/60945/IDL60945.pdf?sequence=2&isAllowed=y>.
- Muhonen, R., P. Benneworth, and J. Olmos-Peñuela. 2020. From productive interactions to impact pathways: Understanding the key dimensions in developing SSH research societal impact. *Research Evaluation* 29 (1): 34–47. <https://doi.org/10.1093/reseval/rvz003>.
- Nwako, Z., T. Grieve, R. Mitchell, J. Paulson, T. Saeed, K. Shanks, and R. Wilder. 2023. Doing harm: The impact of UK’s GCRF cuts on research ethics, partnerships and governance. *Global Social Challenges Journal* 1 (aop): 1–22. <https://doi.org/10.1332/GJSZ3052>.
- Pasanen, T., and I. Barnett. 2019. *Supporting adaptive management: Monitoring and evaluation tools and approaches*. Overseas Development Institute (ODI) working paper 569. London: ODI.
- Patton, M.Q. 2010. *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: Guilford Press.
- Patton, M.Q. 2017. *Principles-Focused Evaluation: The Guide*. New York: Guilford Publications.
- Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. Thousand Oaks: Sage.
- Pinar, M., and T.J. Horne. 2022. Assessing research excellence: Evaluating the Research Excellence Framework. *Research Evaluation* 31 (2): 173–187. <https://doi.org/10.1093/reseval/rvab042>.
- Polonenko, L.M. 2018. The politics of evidence and results in international development: Playing the game to change the rules?, by Rosalind Eyben, Irene Guijt, Chris Roche, and Cathy Shutt. *Canadian Journal of Program Evaluation*. <https://doi.org/10.3138/cjpe.53072>.
- Ramalingam, B. 2013. *Aid on the Edge of Chaos: Rethinking International Cooperation in a Complex World*. Oxford: Oxford University Press.
- Reisman, J., A. Gienapp, and S. Stachowiak. 2007. *A Guide to Measuring Advocacy and Policy*. <https://www.aecf.org/resources/a-guide-to-measuring-advocacy-and-policy>.
- Rogers, P.J., and C.H. Weiss. 2007. Theory-based evaluation: Reflections ten years on: Theory-based evaluation: Past, present, and future. *New Directions for Evaluation* 2007 (114): 63–81.
- Stachowiak, S., J. Lynn, and T. Akey. 2020. Finding the impact: Methods for assessing the contribution of collective impact to systems and population change in a multi-site study. *New Directions for Evaluation* 2020 (165): 29–44.
- Stern, E., N. Stame, J. Mayne, K. Forss, R. Davies, and B. Befani. 2012. *Broadening the Range of Designs and Methods for Impact Evaluations*. Institute for Development Studies. <https://doi.org/10.22163/fteval.2012.100>.
- Temple, L., D. Barret, G. Blundo Canto, M.-H. Dabat, A. Devaux-Spatarakis, G. Faure, E. Hainzelin, S. Mathé, A. Toillier, and B. Triomphe. 2018. Assessing impacts of agricultural research for



- development: A systemic model focusing on outcomes. *Research Evaluation* 27 (2): 157–170. <https://doi.org/10.1093/reseval/rvy005>.
- Thornton, P., T. Schuetz, W. Förch, L. Cramer, D. Abreu, S. Vermeulen, and B. Campbell. 2017. Responding to global change: A theory of change approach to making agricultural research for development outcome-based. *Agricultural Systems* 152: 145–153. <https://doi.org/10.1016/j.agsy.2017.01.005>.
- Ton, G., J. Mayne, J.A. Morell, B. Befani, M. Apgar, and P. O'Flynn. 2019. *Contribution Analysis and the Estimating the Size of Effects: Can We Reconcile the Possible with the Impossible? CDI Practice Paper 20*. Institute of Development Studies.
- UK Research and Innovation. 2021. *UKRI Official Development Assistance Letter 11 March 2021*. Swindon: UK Research and Innovation.
- Vogel, I. 2012. *Review of the Use of 'Theory of Change' in International Development. Review Report*. [Review Report]. UK Department of International Development. <https://www.gov.uk/dfid-research-outputs/review-of-the-use-of-theory-of-change-in-international-development-review-report>.
- Walton, M. 2016. Expert views on applying complexity theory in evaluation: Opportunities and barriers. *Evaluation* 22 (4): 410–423. <https://doi.org/10.1177/1356389016667890>.
- Weiss, C.H. 1997. Theory-based evaluation: Past, present, and future. *New Directions for Evaluation* 1997 (76): 41–55. <https://doi.org/10.1002/ev.1086>.
- Wilkinson, H., D. Hills, A. Penn, and P. Barbrook-Johnson. 2021. Building a system-based theory of change using participatory systems mapping. *Evaluation* 27 (1): 80–101.
- Wilson-Grau, R. 2018. *Outcome Harvesting: Principles, Steps, and Evaluation Applications*. IAP.
- Zacharewicz, T., B. Lepori, E. Reale, and K. Jonkers. 2019. Performance-based research funding in EU Member States—A comparative assessment. *Science and Public Policy* 46 (1): 105–115. <https://doi.org/10.1093/scipol/scy041>.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

