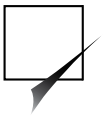

Article

Evaluation
2017, Vol. 23(3) 294–311

DOI: 10.1177/1356389017714382
journals.sagepub.com/home/evi


Evaluating complex interventions: A theory-driven realist-informed approach

Boru Douthwaite
WorldFish, Malaysia

John Mayne
Independent Advisor, Canada

Cynthia McDougall and Rodrigo Paz-Ybarnegaray
WorldFish, Malaysia

Abstract
There is a growing recognition that programs that seek to change people's lives are intervening in complex systems, which puts a particular set of requirements on program monitoring and evaluation. Developing complexity-aware program monitoring and evaluation systems within existing organizations is difficult because they challenge traditional orthodoxy. Little has been written about the practical experience of doing so. This article describes the development of a complexity-aware evaluation approach in the CGIAR Research Program on Aquatic Agricultural Systems. We outline the design and methods used including trend lines, panel data, after action reviews, building and testing theories of change, outcome evidencing and realist synthesis. We identify and describe a set of design principles for developing complexity-aware program monitoring and evaluation. Finally, we discuss important lessons and recommendations for other programs facing similar challenges. These include developing evaluation designs that meet both learning and accountability requirements; making evaluation a part of a program's overall approach to achieving impact; and, ensuring evaluation cumulatively builds useful theory as to how different types of program trigger change in different contexts.

Keywords
agricultural research, case studies, complex systems, realist evaluation, results based management, theory of change

Corresponding author:
Boru Douthwaite, Cushalogurt, Kilmeena, Westport, Co. Mayo, Ireland.
Email: bdouthwaite@gmail.com

Introduction

Many programs seeking to make a difference in people's lives, including programs in the international aid, health, education and agricultural research sectors, are 'complex'. They involve many components and partners, considerable uncertainty in the pathways to impact with numerous feedback loops, and lengthy time frames. There is a growing consensus in the literature that such programs need to be understood as complex interventions in complex systems (Barder and Ramalingam, 2012; Mayne and Stern, 2013; Pawson, 2013). This means that the specific causal links between program intervention and eventual impact are inherently uncertain and emergent (Douthwaite et al., 2003; Patton, 2011). Emergence and uncertainty of outcomes puts a particular set of demands on the management and evaluation of such programs, not least that evaluative practice supports ongoing collective learning so that staff can respond to emerging outcomes of their work and suitably adjust implementation of the program (Loftin, 2014; Patton, 2011; Snowden, 2010; Wild et al., 2015).

In response to this demand, developmental evaluation has emerged over the last six years as an approach to understanding the activities and outcomes of programs operating in dynamic, novel environments with complex interactions (Patton, 2011). Table 1 summarizes the difference between traditional and developmental evaluation.

More recently, the United States Agency for International Development (USAID) coined the term 'complexity-aware' to describe monitoring and evaluation approaches that can deal with uncertainty and unexpected outcomes, including developmental evaluation (Britt, 2013). Complexity-aware approaches are problematic because they challenge orthodoxy in much of mainstream research and evaluation, as the table suggests. However, much of what is written on them is normative and theoretical. Comparatively little has been written on practical experience of developing and using complexity-aware approaches. This is particularly so where learning from experience is most required: in the context of hierarchical organizations with deeply engrained monitoring and evaluation (M&E) practice where structure restricts the nimbleness required for the organization to navigate complexity as well as the space to adopt new approaches that would help to do so.

This article helps fill this gap. We describe the design, development and early implementation of a complexity-aware evaluation approach developed for a program run by a hierarchical and long-established organization – the CGIAR.[1] We present early results, reflect on our experience, develop a set of guiding principles and identify practical implications for other teams wishing to set up and run complexity-aware M&E systems.

The evaluand: The RinD approach

The program for which we developed a complexity-aware evaluation approach was the CGIAR Research Program on Aquatic Agricultural Systems (AAS). The goal of AAS was to improve the wellbeing of poor people dependent on aquatic agricultural systems by putting in place the capacity for communities to pull themselves out of poverty (AAS, 2011). AAS began in 2011 by establishing programs in geographically defined areas called hubs with an aspirational goal to make positive difference on the livelihoods of six million poor and marginalized living in the hubs by 2023 (AAS, 2014). The AAS program established five hub programs by the end of 2013 in Zambia, Bangladesh, the Philippines,

Table 1. Comparison between traditional and developmental evaluation (Patton, 2006: 30).

Traditional evaluations	Complexity-aware, developmental evaluations
Render definitive judgements of success or failure	Provide feedback, generate learning, support direction or affirm changes in direction
Measure success against predetermined goals	Develop new measures and monitoring mechanisms as goals emerge and evolve
Position the evaluator outside to assure independence and objectivity	Position evaluation as an internal, team function integrated into action and ongoing interpretive processes
Design the evaluation based on linear cause-effect logic models	Design the evaluation to capture system dynamics, interdependencies and emerging interconnections
Aim to produce generalizable findings across time and space	Aim to produce context-specific understandings that inform ongoing innovation
Accountability focused on and directed to external authorities and funders	Accountability focused on learning and responding to what is unfolding
Evaluator controls the evaluation and determines the design based on their perspective of what is important	Evaluator collaborates in the change effort to design a process that matches philosophically and organizationally
Evaluation engenders fear of failure	Evaluation feeds hunger for learning

Cambodia and Solomon Islands. The hub programs were divided into smaller units called initiatives.

By 2013, the AAS program developed the research in development (RinD) approach for achieving impact which it implemented and tested in each hub. The RinD approach is described below. The program was clear from the beginning that it was engaging in complex systems reflected in naming its approach 'research-**in**-development' to highlight that its research was embedded in local contexts and evolving development processes. The evaluation approach was developed to learn about and assess the AAS RinD approach.

AAS's overarching program theory was that agricultural research processes (e.g. multi-partner collaborations) and outputs (i.e. new technologies) work to catalyze and foster processes of rural innovation. It is these innovation processes, that may be technical, institutional or both, that lead to development outcomes and impact. AAS developed the RinD approach to build the capacity of hub innovation systems to innovate faster and more equitably in favour of the poor and marginalized. The RinD approach does so by building research collaborations across institutional and scale boundaries (e.g. between farmers and researchers, or between different government ministries). This program theory was unorthodox within the CGIAR: all other CGIAR Research Programs (CRPs) build their program theory around the adoption and use of new technology. RinD was not a rigid framework but instead evolved through on-going learning from practice adapted to context in each hub.

The RinD evaluation approach

History

When AAS began in 2011 the emphasis was to establish the program in the five geographic hubs and to develop the RinD approach. The program's Knowledge Sharing and Learning (KS&L) Theme was responsible for program M&E and developed a framework for staging the M&E

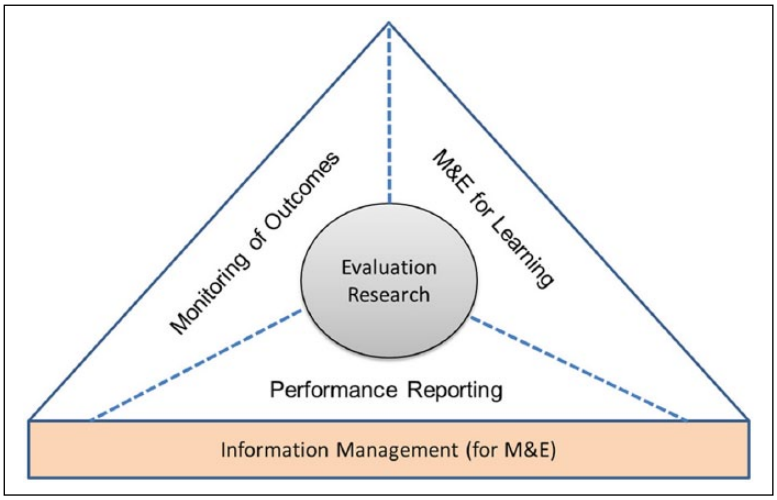

Figure 1. Framework for staging the AAS M&E system (beginning at the bottom of the triangle). From Douthwaite et al. (2014: 7).

system, shown in Figure 1. Foundational work on the M&E system began by setting up an information management system in support of performance reporting so as to meet basic program accountability requirements to donors. The next step was to implement the other three parts of the framework: to monitor outcomes, and to build M&E for learning and evaluation research. Monitoring of outcomes were the methods to be used to track program progress towards its impact goals. M&E for learning were methods in support of reflexive practice, in particular building and revisiting location-specific theories of change as participatory action research. Evaluation research were the capstone activities by which the program would build middle-range theory[2] useful for other programs attempting similar work. At this point, program staff were staging the development of the M&E system assuming the program would last 12 years.

AAS's M&E strategy was built on a report that the program commissioned to make the case for using theory-based evaluation for programs such as AAS (Mayne and Stern, 2013). AAS commissioned the report in anticipation of challenge from the CGIAR with respect to program M&E.

The challenge arrived earlier than expected. The program submitted an extension proposal in 2014 to the CGIAR's oversight organization, the Independent Science and Partnership Council (ISPC). In their review of the proposal, the ISPC were critical of the RinD approach saying it was 'an excessive shift away from bio-technical innovation research toward an experiment in development process' (ISPC, 2014: 1). The ISPC was also critical of the program's evaluation approach, calling for the use of counterfactuals. In response, the program stressed its use of theory-driven evaluation and pointed to practical and ethical issues with using control groups. The response highlighted that the program was addressing a comparative and overarching research question to address the counterfactual issue and guide the design of the top part of the triangle. The question was:

> How, and in what situations, does the AAS Research in Development (RinD) approach foster enduring and equitable change in livelihoods of the poor and marginalized in aquatic agricultural systems – and how are these changes different from those produced by other approaches?

The program engaged with ISPC to find mutually-agreeable methods to evaluate AAS performance, but with little success. Between 2014 and October 2015 the CGIAR received a series of cuts amounting to 33 per cent of core funding to CRPs. The Consortium Office, who work on behalf of the CGIAR research centers, concluded that funding cuts should be decided based on performance ranking of (CGIAR Consortium, 2015). AAS achieved a failing grade on the basis of the ISPC evaluation of its extension proposal and was closed in 2016 along with the two other CGIAR system CRPs.

Design

Overall design. AAS researchers developed the RinD Evaluation Approach to tackle our comparative research question. Our starting point was the requirement stated in the M&E strategy that evaluation should help understand how RinD is working in context, fast enough to inform improvements in program implementation. This suggested taking a theory-driven approach.

We then developed a theory of change to describe how RinD works in a geographic hub (see Figure 2). The numbers in the following narrative refer to outcome boxes in the ToC.

The ToC is a model based on a mixture of early evidence, stakeholder theory and existing literature. The RinD approach starts by identifying a commonly-agreed hub development challenge to provide a focus for engagement. For example, the hub development challenge in Zambia was 'to make more effective use of the seasonal flooding and natural resources of in the Barotse flood plain system'. Community-level engagement is through participatory action research (PAR) that involves facilitating the creation of visions of success with respect to tackling the challenge and implementing and revisiting actions plans to achieve them. Hub-level engagement involved agreeing and implementing a set of research initiatives to address opportunities emerging from the community-level research as well as to pursue opportunities identified at other scales, e.g. modelling water flow within the Barotse flood plain system so as to be better able to predict flooding.

The main result of implementing RinD is the creation of safe spaces for experimentation, action, reflection, questioning and learning for those involved (1). Safe spaces can take several forms including group meetings, workshops and after action reviews at community, initiative and hub scale. The ToC assumes that working in safe spaces through PAR leads to the generation of research output including technology and knowledge (2), increases in social capital and collective efficacy (3) and increases in understanding about how change happens and how to trigger it through building and revisiting theories of change (4). These three results directly build the capacity of hub actors to innovate. Working in safe spaces using a gender transformative approach leads to changes in norms and socially defined roles (5) that in turn leads to more gender and socially-equitable control of assets and decision-making (6). More equitable control of assets and decision-making by hub actors influences the participation, interactions and decision-making that takes place as part of hub innovation processes. Hence this outcome builds system capacity to innovate more equitably (7) that results in faster and more equitable innovation processes (8) as the pathway to improve the livelihoods of the poor and marginalized (9). The ToC is built on a number of assumptions, implicit in the linking arrows. Critical assumptions are made explicit in the diagram, particularly those to do with the need for quality of process and time to implement and these are tested and revised in subsequent iterations of revisited ToCs.

There are different theory-driven approaches to evaluation. The RinD Evaluation Approach is based on a realist one (Pawson, 2013; Pawson and Tilley, 1997; Westhorp, 2014) because of

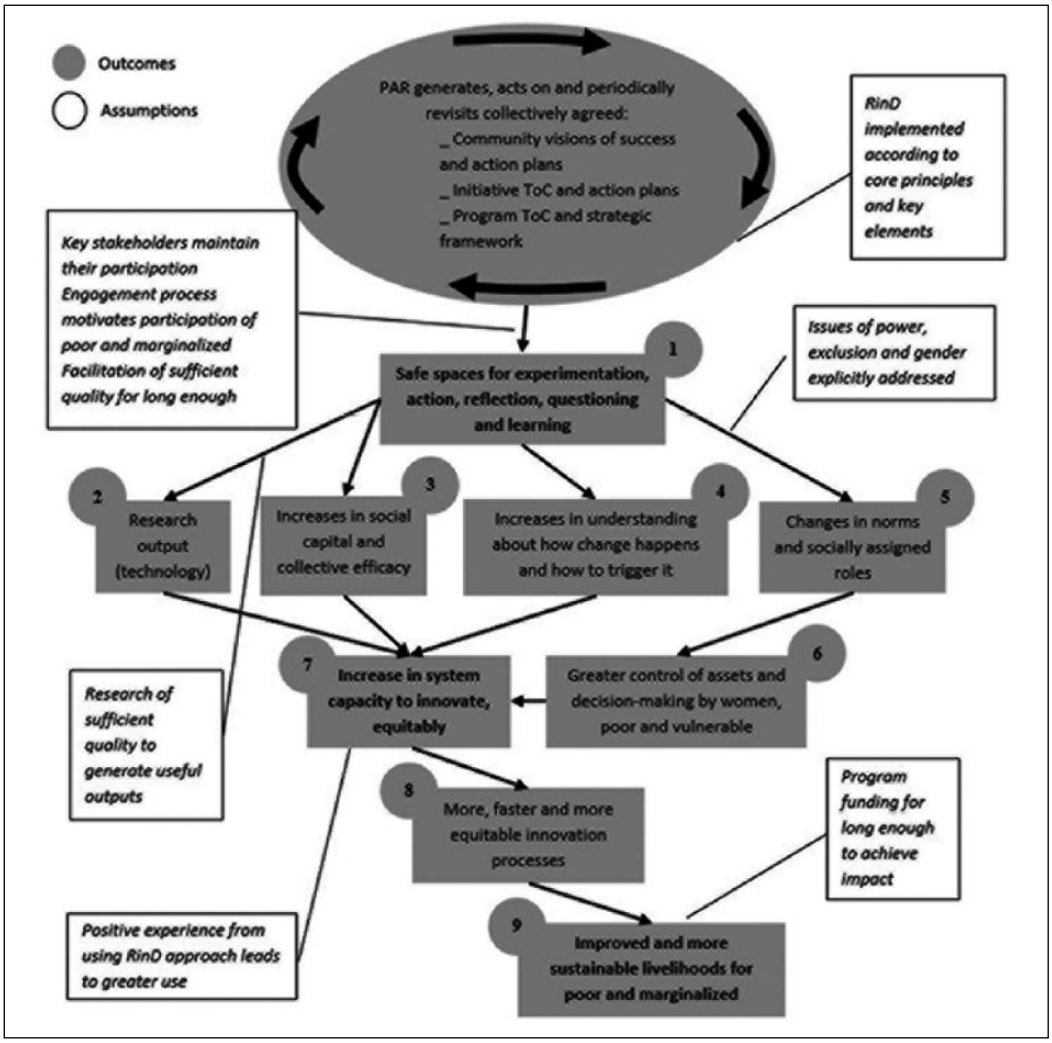

Figure 2. ToC for how RinD works in a hub.

its emphasis on building and testing causal hypotheses, something that is familiar to researchers working in the CGIAR and thus, we thought, more likely to be understood and accepted.

Certain aspects of realist evaluation were particularly influential in designing the RinD evaluation approach, including:

- The focus on identifying underlying causal mechanisms and how the program triggers them (see Box 1).
- The Popperian idea that there is nothing absolute about theory; that it builds on foundations that are firm enough (Popper, 2005). Theories are born, evolve and are superseded. From this perspective, the task of evaluation research is to build causal theory from literature and stakeholder experience and probe into the parts of it that rest on shaky ground.

- The realist evaluation question – How does an intervention work, what aspects, to what extent, for whom and in what contexts?
- The premise that there are no silver bullets – nothing works everywhere all the time.
- The idea that programs don't change people, it is how people interpret and use what the program provides that changes things.
- The idea that evaluations should explicitly contribute to broader theory-building so that learning from evaluation accumulates over time and can cross sectoral boundaries, e.g. that learning from works or not in the health sector can inform agriculture and vice versa (Pawson, 2013).

Box 1. Examples of causal mechanisms.

The concept of causal mechanisms is fundamental to the generative view of causality, and realist evaluation in particular, but is also a cause of misunderstanding (Westhorp, 2014). Gravity is an example of a causal mechanism in the physical world. Gravity is what causes an apple to fall from my hand to the ground. Whether the apple falls or not depends on whether I release my grip. Letting go of the apple is the trigger. Social norms are an example of a mechanism in the social world (Elster, 2007). Social norms suggest a certain way of acting in particular circumstances. For example, whether I act in accordance to the expected behavior of not talking on my mobile in a train carriage will depend on triggers such as a disapproving glance from a fellow passenger or a sign asking passengers to respect others' wish for quiet. The outcome of triggering a mechanism depends on context. If I release an apple at the bottom of a swimming pool it will float because buoyancy replaces gravity as the dominant mechanism. Whether I make a phone call in the railway carriage will depend on the urgency of the situation. Both gravity and social norms are real, but their working is not directly observable. The 'under the surface' nature of mechanisms is a fundamental characteristic.

Applying this realist theory-based perspective in the RinD evaluation approach means that the RinD evaluation design focuses on building RinD's ToC (Figure 2) and testing the parts of it that are novel, or appear to rest on theoretically or empirically shaky ground. As hubs implement RinD, the M&E system seeks empirical evidence to support, contradict or modify more detailed sub-theories, also known as 'nested' theories of change that sit within the overall RinD ToC. The result of this analysis combines theoretical understanding and empirical evidence, and focuses on explaining the relationship between the context in which RinD is applied, the mechanisms by which it works, or doesn't, and the outcomes which are produced. Periodically the overall RinD ToC is revisited through a realist review process informed by the hub-specific work.

The RinD evaluation design is comparative as well as theory-based. In particular, it responds to the overarching research question of how RinD compares against other research for development (R4D) approaches. The AAS Program offered a natural experiment by also working with projects that use R4D approaches focused more on developing and scaling out technology than building capacity to innovate and adapt. The evaluation approach made use of this. Like RinD, R4D approaches are also complex interventions acting in complex systems. Hence the approach is the same - to build and test theories that explain how R4D approaches are expected to work and the outcomes they are expected to produce.

Additionally, the RinD evaluation approach uses case study methodology (Yin, 2014), including single, multiple and nested case studies. The approach is 'mixed methods' in nature,

combining qualitative and quantitative data analysis and the use of multiple data gathering and analytical methods. This includes PAR documentation, longitudinal panels and statistical analysis of data from trend lines. No one method will necessarily answer the research questions of interest. As Stame (2004: 60) writes: 'All methods can have merit when one puts the theories that can explain a program at the center of the evaluation design. No method is seen as the "gold standard"'. For example, proving the benefits of an innovation developed through RinD might best be done using a randomized experimental design, particularly if the innovation is relatively simple and its benefits potentially large, thus justifying the investment in such a design.[3]

Research on the RinD approach took place primarily in the AAS hubs, in a set of activities called the RinD initiative. Research findings from each hub were further analyzed across hubs using case study and synthesis approaches.

Detailed design. The evaluation approach addressed five sub questions to be answered over time, building a body of evidence and theory in the process.

1. What are the underlying development trends where the program works?
2. Whether, for whom and how are different aspects[4] of RinD and R4D approaches working, and in what contexts?
3. To what extent and at what scale is RinD and R4D working?
4. How do RinD outcomes compare to those achieved by other R4D approaches?
5. What are the key challenges to implementing RinD and how can they be overcome?

The approach was tailored to each hub. The methods used to answer the sub questions in Bangladesh are explained below.

1: What are the underlying development trends where the program works?

The evaluation system must be able to make credible causal claims, if the empirical evidence and theory developed to explain how AAS research works is to be credible and useful. The research must discount the explanation that the changes would have happened anyway. Hence the evaluation approach intended to track progress in RinD- and R4D-focal villages against key developmental outcome indicators for participating and non-participating households.

2: Whether, for whom and how are aspects of RinD and R4D approaches working, and in what contexts?

The overall approach to answering this question was to develop hypotheses relating to whether and how aspects of RinD and bilateral R4D approaches lead to outcomes, and test them. Differences between outcome trends for participating and non-participating households suggest hypotheses. The differences may or may not[5] be the result of RinD or R4D intervention. Whether a causal claim is valid – whether the hypothesis is proven – will depend on identifying and confirming it was the RinD or R4D approach that triggered the causal mechanism(s) that accounts for the difference. The causal mechanisms were to be identified using outcome evidencing (Paz-Ybarnegaray and Douthwaite, 2016), a method developed by the AAS program based on outcome harvesting (Wilson-Grau and Britt, 2012). A smaller sub-set of households were to be selected to serve as a **longitudinal panel** to identify and provide qualitative explanation of causal mechanisms. Follow-up interviews were to be carried out as necessary.

3: To what extent and at what scale is RinD and R4D working?

This research question focused on understanding for whom RinD is working and to what extent its use and outcomes have spread. This was done through analysis of program evidence streams that identify adoption, including from outcome harvesting, project M&E and PAR documentation for different groups. PAR documentation involved codifying who is participating in RinD activities and who is not and identifies outcomes that can likely be attributed to participating in RinD.

Further bespoke **adoption surveys** were foreseen to quantify the spread and effects of RinD and R4D outcomes. Part of answering the research question was to build scaling theories of change, i.e. causal descriptions of how program outcomes are scaling out to others.

4: How do RinD outcomes compare to those achieved by other R4D approaches?

We started to answer this question using realist review (Pawson et al., 2005). The first step was to build overall theories of change describing RinD and the specific R4D approaches chosen for comparison in each hub. The overall ToC for RinD is shown in Figure 2. The approach pulled together empirical evidence and results of theory-building and testing across hubs to add detail to the RinD and R4D theories of change, supporting, contradicting or modifying the theories as it goes. This synthesis was to have been repeated periodically. RinD was to be compared to other R4D approaches by comparing and contrasting the respective theories of change. For example, the analysis might have found that while all agricultural research builds capacity to innovate, different approaches build different dimensions of it; that RinD works on actors' motivations, linkages and decision-making while more technology-focused approaches increase the stock of novelty available to trigger new innovation trajectories. This analysis allows for the generation of middle-range theory[6] that explains how families of agricultural research approaches work to produce different types of outcomes, for different types of beneficiaries in different contexts. This theory, and its use by the people who fund, plan and implement agricultural research programs, has potential to improve practice.

5: What are the key challenges to implementing RinD and R4D and how can they be overcome?

This question is answered through an **annual after action review** at which program staff and key partners reflect on what worked, what didn't work and what to change for the following year. The question is important because RinD and R4D outcomes depend crucially on quality of implementation which in turn depends on successfully tackling challenges to implementation.

Each hub produces an **annual evaluation research report** each year to provide answers to the five questions. The answers build on those of the previous year. Implications feed into the annual planning cycle. The reports are the key input into the realist review described above.

Early implementation

In early 2016, the CGIAR closed the AAS program as described above, however not before program staff had begun implementing four areas of work: after action reviews at hub and program level, development and revisiting of theories of change; development of the outcome evidencing method; clarification of the modus operandi of the RinD approach; and, establishment of trend lines. We briefly review progress and results for each.

Hub staff carried out annual after action reviews from the beginning of the program by reflecting amongst themselves and with key stakeholders on what had worked well, not so well and what to change for the following year. In 2014, hub staff were asked by the program team to reflect on how the RinD approach was working and the outcomes it was starting to produce in their respective hubs. The program then brought key staff involved in the hub reflections to headquarters in January 2015 to distill out cross-hub learning of use both for hub teams and other research for development programs. The workshop led to the publication of an AAS Working Paper in which four areas were explored in detail (Douthwaite et al., 2015): community engagement, partnerships, integration of gender transformative approaches into RinD and learning how to make science more inclusive. A number of insights resulted from this work including, for example, a clearer articulation of the value of RinD and a better understanding of how it was working. Staff involved became clearer on their own multiple roles as researches, facilitators and knowledge brokers. The lead authors further used the learning to reflect on how RinD compares to their previous experience working with conventional research for development approaches and what it takes to institutionalize a 'new professionalism' required to implement RinD. This paper was published in the International Journal of Agricultural Sustainability (Douthwaite et al., 2017). The paper concluded that while possible to take a complexity-aware approach in a hierarchical organization, caution is required to ensure there is the time, space and appropriate evaluation methodologies in place to appreciate outcomes different than what conventional agricultural research aspires to.

Program staff supported hubs to develop theories of change from the outset, beginning with aspirational models developed with key hub stakeholders as a way of agreeing a common hub development challenge and the opportunities for tackling it. Program staff also supported the development of theories of change of research initiatives developed to exploit the opportunity, and their revisiting through a PAR process. Staff involved in Solomon Islands, Zambia and at headquarters reflected on their experience resulting in a paper to be published in the Action Research Journal (Apgar et al., 2017). They concluded that the power of working with theory of change was in incorporating stakeholder and real world findings in to the research process, but this depended on: donors and staff better appreciating emerging outcomes; building the capacity of stakeholders to reflect more critically. They concluded that the individuals responsible for designing the AAS evaluation system must also be involved in implementing it on the ground given it is not straightforward to make it work in practice across different contexts.

Program staff working in all five hubs developed the outcome evidencing approach to identify and make sense of emerging program outcomes, both expected and unexpected. A paper describing the approach was published in the American Journal of Evaluation (Paz-Ybarnegaray and Douthwaite, 2016). In the approach, staff and change agents on the ground identify outcomes resulting from program intervention in each hub, making sense and validating them. This involves clustering outcomes together and describing causal linkages within the clusters using a multi-cause diagram. The clusters were similar across hubs, supporting the idea (Scriven, 1976) that successful approaches have a particular modus operandi. Paz-Ybarnegaray and Douthwaite (2016) describe this modus operandi and explore how systems concepts such as catalytic probes, attractors, beneficial coherence, emergence and strategic niche management apply to AAS and might be used to explain how RinD appears to be working. Some early outcomes were also described, including how the program in the Philippines successfully sought official recognition from the Regional Development Council in the Philippines, which is the high-level policy-making body that serves as the regional counterpart for the National Economic and Development Authority (NEDA).

In an article for Agricultural Systems, Douthwaite and Hoffecker (2017) looked in depth at success stories to identify the modus operandi of the RinD approach. They identified five causal elements that the RinD approach provided in both cases:

1. A process to engage stakeholders in developing a joint vision of success
2. A process to identify an issue of common interest
3. Facilitation of engagement between existing stakeholders and linkages to new stakeholders
4. 'Safe space' for stakeholders to build trust and develop working relationships
5. Opportunities to 'learn by doing' supported by coaching
6. Knowledge inputs with high relevance to local stakeholders.

The fourth piece of work was to start to establish trend lines in each of the hubs. The work began by identifying the main outcomes that AAS wished to track, with the idea that it would be possible to agree on a small set of 'bellwether' indicators that would indicate broader trends and reduce data gathering. This proved impossible because the six theme leaders involved in the selection wanted to see their key work represented by an indicator, fearing that otherwise it be deemed less important and be put at risk. The eventual survey instrument ran to 34 pages and when enumerated in Zambia required a team of 10 AAS employees to work for around 90 days and cost $90,000 including basic data analysis. Data was collected in both AAS and partner focal villages to establish a base line for the selected indicators, with the intention to repeat every two years so as to establish trends. Had the trend line been repeated in other hubs, ways would have had to be found to reduce the cost in terms of people's time and cost.

Discussion

The experience of designing and implementing the RinD evaluation approach, together with existing literature, allows us identify six design principles of potential use to other programs wishing to design a complexity-aware M&E system.

Stage the roll-out of M&E system

We learned that the M&E system needs to evolve with the program. The first priority is to put the monitoring and information management systems in place to meet basic accountability requirements. The boundaries between monitoring and evaluation are fuzzy. Regular after action reviews worked well to establish the practice of reflexive learning. After that, building the capacity and organizational structure to meet other requirements of the system should be staged according to when the program needs it. For example, the ability to monitor emerging outcome pathways is needed when change starts to happen on the ground. Monitoring change before the outcome pathways become clear can suck up a lot of resources that could be better used to foster the change in the first place.

Contribute to achieving the program's overall goals

Programs operating in complex settings require an M&E system to help them navigate that complexity. M&E efforts should be part of how the program is implemented and managed to

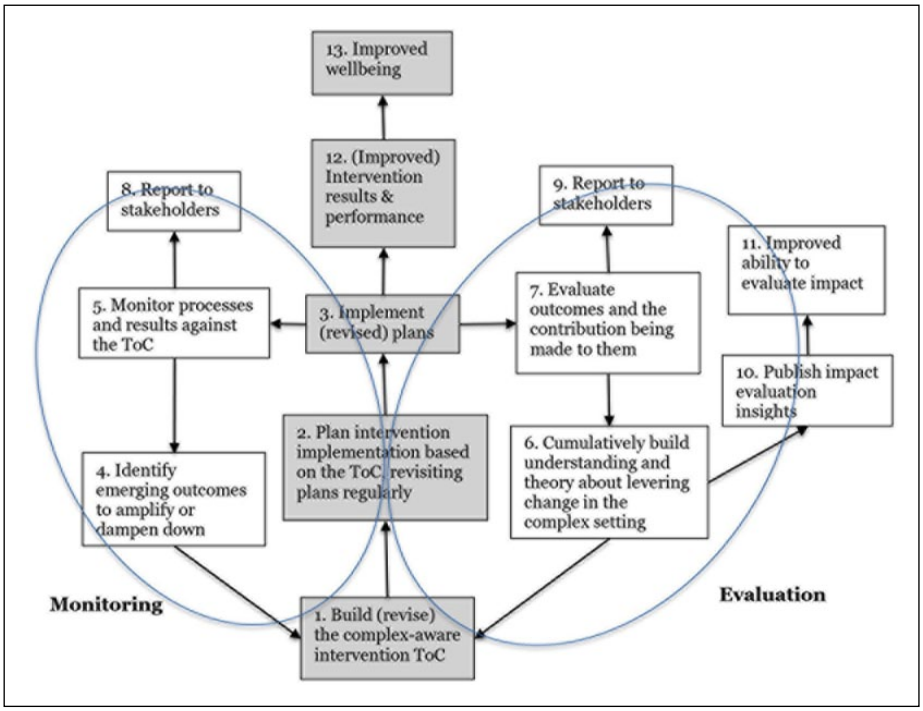

Figure 3. A theory of change for M&E.

meet its goals; not as an oversight or audit function, but as a mainstay of the program's ability to learn, identify threats and opportunities and adapt accordingly. The RinD evaluation approach addressed this principle in several ways. Outcome harvesting and after action reviews identified early patterns of change resulting from program intervention and insight on how to support them. Understanding how different types of research approach work in different contexts helped staff clarify the added value of RinD compared to conventional approaches and be clearer on the skill set required to realize the potential.

Putting it another way, complexity-aware monitoring and evaluation should operate its own theory of change that makes it clear how it contributes to program learning and goals. Figure 3 shows one that we were beginning to see work in AAS. The numbers in the following narrative refer to the boxes in the figure.

A program builds and revises its theory of change (1) for the overall program and for specific interventions nested within it. Theory of change is used to plan and implement program activity (2 & 3). The left side of the figure (4 & 5) shows how regular monitoring identifies emerging outcomes that can be further supported while those not leading to positive change are dampened down. This leads to a revised ToC (1) with subsequent adjustments to the implementation (2 &3), and the cycle repeated. The right side of the figure shows evaluation playing a similar role (6 & 7), on a periodic basis. Evaluation research works with middle-range theory that allows learning to cumulate and be useful for similar types of program and intervention (7). This learning leads to a more informed ToC (1) and further revisions to how the intervention is implemented. The insight gained on evaluating impact can then be published for the wider evaluation community (8).

Support learning for adaptive managing that feeds back into the annual planning cycle

The M&E system should produce insight and learning fast enough to help the program adapt to emerging opportunities, threats and unforeseen circumstances as they happen. The idea of quick use of evaluation findings is consistent with Developmental Evaluation (Patton, 2011).

In practice, the RinD evaluation approach was built on a learning process – the building and testing of theories of change during implementation and regular after action reviews – much of which is done collectively as part of participatory action research. This is particularly useful for programs that operate in multiple locations, as emerging outcome pathways in one site may be applicable in another.

Support program accountability requirements

M&E should provide donors and other stakeholders with information on progress towards agreed goals. Unpredictability and emergence means that the actual pathways towards the goals, and indeed the scale and nature of the goals themselves, may change over time. This represents a significant challenge to meeting accountability needs since often accountability is seen as meeting key specific milestone targets. There is a need for a different perspective on accountability, namely a focus on being accountable for learning, for know how well the intervention is unfolding (its ToC) and contributing to observed outcomes, and being able to report on key outcomes achieved (Mayne 2007).

In practice, the evaluation design provided for robust accountability reporting on:

- *Learning*. The learning that has taken place—the improvements in implementation that have been made based on empirical information on what is working and what is not working
- *Evolving pathways*. The changes in pathways and theories of change that have occurred based on better understanding of the complex setting
- *Progress along pathways*. The observed outcomes and impacts along the pathways that have occurred to date
- *Deviations*. Explanation of any deviations from prior targets
- *Likelihood of future impacts*. Estimations of the likelihood of meeting future targets based on current evidence and understanding
- *Improvements of targets*. Changes made to any future targets and the justification

Be implementable by staff and the available budget

This principle is perhaps the most important. Whatever M&E system that is built must be practicable in terms of budget and the program's ability to implement. Interventions in complex settings require staff to be much more engaged in measuring and analyzing than might otherwise be the case. There is also the need to be flexible, adapting and responding to conditions and events as they unfold over time and greater understanding is gained.

In practice, the RinD evaluation approach describes a body of research that is well within the means of CGIAR centers and partners to implement. The requirement for evaluation research that cumulates evidence and builds a body of theory over time implies a greater contribution of social science capacity to evaluation than has been normal practice in the

CGIAR and perhaps in other development interventions. It implies a blurring of the traditional boundaries separating 'M&E' from 'Impact Assessment' from 'Social Sciences' which has to be negotiated. Trend line work can be very expensive and beyond some hubs' budgets to implement.

We learned that any evaluation approach of a complex program is itself likely to be complex. Implementing the RinD evaluation approach in practice required adaptation to local context, capacities and budget and was not always straightforward. The need for capacity development should not be underestimated.

Contribute to the development of useful impact evaluation methods

Practical methods for impact evaluation in complex-aware settings is an ongoing issue in evaluation (Stern et al., 2012). Often more traditional counterfactual approaches are not appropriate nor feasible. A reasonable expectation then is that new approaches that are useful be communicated to the evaluation community. This was especially the case with the AAS program where the CGIAR had a solid background in using experimental approaches in assessing impact of its efforts.

In practice, the contribution the RinD approach makes within the sometimes traditional CGIAR system is that it is able to understand how different types of research for development interventions are working fast enough to influence on-going implementation. It is able to work without the sometimes-problematic need for a counterfactual by identifying underlying mechanisms that are causing outcomes and establishing program contribution 'beyond reasonable doubt'.

Implications for evaluation in complex systems

Based on experience to date, six implications are identified for the design of evaluation approaches in other programs who engage in complex systems.

Think of evaluation as an integral part of the program's M&E system

Often evaluation is thought of as a series of one-off studies each with their own terms of reference (Pawson, 2013). This paper has described an evaluation approach that is an integral part of the program M&E system that uses research process to systematically establish and test assumptions in order to adapt and learn their way towards impact. Such approaches are increasingly seen as essential to meet accountability requirements in complex and adaptive settings (Douthwaite et al., 2004; Earl et al., 2001; Patton, 2011).

Develop a ToC for program M&E

It may help make the case for investment in program monitoring and evaluation to have a specific ToC that shows how M&E contributes to program learning and goals, and to verify and revise the ToC as implementation proceeds. Theories of evaluation are discussed in the literature. A recent example is found in *Evaluation and Program Planning* Special Section in the 2013 Volume 38 on 'Using Logic Models to Facilitate Comparisons of Evaluation Theory'.

See evaluation contributing to a body of program theory and make that case

As the RinD ToC makes clear, we agree with Weiss (1997) that programs are theories made incarnate. Our argument is that given this, programs, and particularly research programs, should test this theory as part of implementation, as a strategy for learning their way towards impact in complex systems. This view is supported by Pawson (2013: 86) who argues that evaluation findings should contribute to a body of theory over time, which he calls 'recyclable conceptual frameworks'. 'Rather than starting each inquiry from scratch, a stock of recyclable conceptual frameworks is created to distinguish different classes of interventions and to set out their component theories. All evaluations then operate within a common set of program theories [theories of change], each inquiry being capable of adding to and refining that framework.'

Make the case for complex theories of change

Complexity-aware evaluation has the opportunity to build empirical evidence and understanding as to how emergence and positive feedback loops in interventions can be triggered in practice. In this way complex theories of change can show more plausible pathways by which relatively small development investments can lead to large-scale impact. Part of doing this in practice involves identifying critical parts of an overall program ToC where non-linear responses are expected and then developing more detailed 'nested' theories of change for these. This is also a strategy for disaggregating complex ToC into manageable parts – nested theories of change (Mayne, 2015).

Consider using elements of the RinD evaluation approach

AAS has put thought and effort into designing an evaluation approach that is able to meet a set of oftentimes conflicting design criteria. There may be elements of the design approach and design itself that are useful to other programs.

Provide for a broader perspective on accountability in complex settings

Accountability in complex settings not only needs to show progress along impact pathways and increasing confidence in the likelihood of achieving future impact, but also that the program is using M&E data to gain a better understanding of the system it is trying to change, and taking management decisions based on this understanding. This is in line with a call made by Earl et al. (2001) and the Outcome Mapping movement for make recipients accountable for demonstrating that they are progressing towards impact and improving effectiveness, not for developmental impact itself, which in any case nearly always occurs well after a project has finished.

Clarify and engage with broader system expectations

Like AAS, most programs operate as part of a broader system upon which they depend in part for legitimacy, funding and recognition. How the broader system views the program is linked to the information the program provides that in turn is linked to the program's M&E system. In this context, choice of evaluation methods and use of evaluation findings may lead to tension

between the program and the broader system, especially when the program is engaging in new and innovative approaches to deal with complex settings. Advocates of new approaches must recognize the need to explain them, prove their worth and show where they fit into the broader system. On reflection, AAS should have been clearer earlier with respect to its theory of change, basis for causal inference and design principles underpinning its M&E system. However, in some organizations complexity-aware M&E may never be accepted given its challenge to orthodox approaches and with whom the power to withdraw funding rests.

Concluding remarks

Increasingly programs that aim to bring improvements to people's lives are being understood as complex interventions in complex systems. With this realization comes a growing need for monitoring and evaluation approaches that can unpick and explain program contribution to on-going processes of change so that donors can see the benefit of their investment and program staff can learn and improve. This paper has described the RinD evaluation approach designed to be able to meet this dual accountability and learning function. It was designed to tease apart the many causal factors at work in the complex settings within and across sites to establish the extent to which program intervention made a difference and the extent of that change. The approach seeks to identify the interplay between triggers, mechanisms and context to identify 'portable' learning and to build and test evidence-based theory. The learning, evidence and theory are expected to be useful to inform investment, design and implementation of agricultural research for development projects and programs. Evaluations of complex programs are likely themselves to be complex, requiring adaptation to local context and budget and implementation was not always straightforward. The need for capacity development of those implementing should not be underestimated.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. The CGIAR is a worldwide partnership, founded in 1971, addressing agricultural research for development carried out by 15 research centers in Africa, Asia, Europe and the Americas. As of 2014, the CGIAR employed more than 8500 researchers and support staff worldwide, with an annual budget of US$800 million (Agropolis International, 2015) controlled through a central Consortium Organization.
2. Middle-range theories are developed at a middle level of abstraction between published theory and action on the ground so that they can be used to guide the implementation and evaluation of similar families of intervention (Pawson and Tilley, 1997).
3. See Poverty Action Lab for discussion on when randomized experiments are appropriate, and not appropriate. http://www.povertyactionlab.org/methodology/when/when-randomization-not-appropriate
4. The different RinD or R4D principles or methods, for example use of PAR, role of critical reflection, use of theory of change, gender transformative approaches, etc.
5. The assumption is that it is impossible to select households and control conditions to rule out differences occurring due to causes other than participation in AAS.
6. Middle-range theories are a realist concept. They are positioned between universal social laws on one hand and contextual stakeholder theory on the other (Pawson, 2013).

References

Apgar M, Allen W, Albert J, et al. (2017) Getting beneath the surface in program planning, monitoring and evaluation: Learning from use of participatory action research and theory of change in the CGIAR Research Program on Aquatic Agricultural Systems. *Action Research Journal* 15(1): 15–34.

[AAS] CGIAR Research Program on Aquatic Agricultural Systems (2011) Program proposal. Available at: pubs.iclarm.net/resource_centre/WF_2936.pdf (accessed 2 July 2015).

[AAS] CGIAR Research Program on Aquatic Agricultural Systems (2014) Extension proposal 2015-2016. Available at: https://goo.gl/8hQT8I (accessed 8 August 2016).

Agropolis International (2015) CGIAR Consortium headquarters in Montpellier. Available at: www.agropolis.org/cooperation/headquarters-cgiar-consortium.php (accessed 11 September 2015).

Barder O and Ramalingam B (2012) Complexity, adaptation, and results. *Global Development: Views from the Center*. Available at: blogs.cgdev.org/globaldevelopment/2012/09/complexity-and-results.php (accessed 11 September 2015).

Britt H (2013) US Agency for International Development (USAID), Bureau for Policy, Planning and Learning. *Discussion note: Complexity aware monitoring*. Available at: https://usaidlearninglab.org/library/complexity-aware-monitoring-discussion-note-brief (accessed February 2017).

CGIAR Consortium (2015) CGIAR Strategy and Results Framework 2016-2030. Available at: www.cgiar.org/resources/strategy-and-results-framework/ (accessed September 2015).

Douthwaite B and Hoffecker E (2017) Towards a complexity-aware theory of change for participatory research programs working within agricultural innovation systems. *Agricultural Systems* 155: 88–102.

Douthwaite B, Apgar M and Crissman C (2014) Monitoring and Evaluations Strategy Brief. Penang, Malaysia: CGIAR Research Program on Aquatic Agricultural Systems. Program Brief: AAS-2014-04.

Douthwaite B, Apgar M, Schwarz A, et al. (2015) Research in development: Learning from the CGIAR Research Program on Aquatic Agricultural Systems. Penang: CGIAR Research Program on Aquatic and Agricultural Systems. Working Paper. AAS-2015-16.

Douthwaite B, Apgar JM, Schwarz AM, et al. (2017) A new professionalism for agricultural research for development. *International Journal of Sustainable Agriculture*.

Douthwaite B, Ekboir JM, Twomlow S, et al. (2004) The concept of integrated natural resource management (INRM) and its implications for developing evaluation methods. In Shiferaw B, Freeman HA and Swinton SM (eds) *Natural Resource Management in Agriculture: Methods for Assessing Economic and Environmental Impacts*. Wallingford: CABI Publishing, 321–40.

Douthwaite B, Kuby T, Van De Fliert E, et al. (2003) Impact pathway evaluation: An approach for achieving and attributing impact in complex systems. *Agricultural Systems* 78(2): 243–65.

Earl S, Carden F and Smutylo T (2001) *Outcome Mapping: Building Learning and Reflection into Development Programs*. Ottowa, Canada: International Development Research Centre (IDRC).

Elster J (2007) *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

ISPC (2014) Draft ISPC Comments on the revised extension proposal of the CRP Aquatic Agricultural systems (AAS) for 2015–2016. Available at: https://goo.gl/KaBX6A (accessed 31 July 2016).

Loftin MK (2014) Truths and governance for adaptive management. *Ecology and Society* 19(2): 21.

Mayne J (2007) Evaluation for accountability: Reality or myth? In Bemelmans-Videc M-L, Lonsdale J and Perrin B (eds) *Making Accountability Work: Dilemmas for Evaluation and for Audit*. New Brunswick, NJ: Transaction Publishers, 63–84.

Mayne J (2015) Useful theory of change models. *Canadian Journal of Program Evaluation* 30(2): 119–42.

Mayne J and Stern E (2013) Impact evaluation of natural resource management research programs: A broader view. *ACIAR* 84: 79.

Patton MQ (2006) Evaluation for the way we work. *Nonprofit Quaterly Spring*: 28–33.

Patton MQ (2011) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: Guilford Press.

Pawson R (2013) *The Science of Evaluation: A Realist Manifesto*. London: SAGE.
Pawson R and Tilley N (1997) *Realistic Evaluation*. London: SAGE.
Pawson R, Greenhalgh T, Harvey G, et al. (2005) Realist review–a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research &Policy* 10(1): 21–34.
Paz-Ybarnegaray R and Douthwaite B (2016) Outcome evidencing: A method for enabling and evaluating program interventions in complex systems. *American Journal of Evaluation* 38(2): 275–93.
Popper K (2005) *The Logic of Scientific Discovery*. London: Routledge Classics.
Scriven M (1976) Maximizing the power of causal investigations: The modus operandi method. *Evaluation Studies Review Annual* 1: 101–18.
Snowden D (2010) *Informed by Knowledge: Expert Performance in Complex Situations*. New York: Psychology Press Ltd, 223–34.
Stame N (2004) Theory-based evaluation and types of complexity. *Evaluation* 10(1): 58–76.
Stern E, Stame N, Mayne J, et al. (2012) Broadening the range of designs and methods for impact evaluations. *Report of a study commissioned by UK Department of International Development*.
Weiss CH (1997) How can theory-based evaluation make greater headway? *Evaluation Review* 21(4): 501–24.
Westhorp G (2014) *Realist Impact Evaluation: An Introduction*. London: ODI Annual Reports.
Wild L, Booth D, Cumming C, et al. (2015) *Adapting Development: Improving Services to the Poor*. London: ODI Annual Reports.
Wilson-Grau R and Britt H (2012) Outcome harvesting. Ford Foundation. Available at: http://www.outcomemapping.ca/resource/outcome-harvesting (accessed June 2012).
Yin RK (2014) *Case Study Research: Design and Methods*, 5th edn. Thousand Oaks CA: SAGE.

Boru Douthwaite is an agricultural engineer, technology policy analyst and independent evaluator who has 25 years' experience working in Africa, Asia and Latin America. His research is aimed at understanding how research output and process can be used to catalyse and bolster rural innovation processes, in particular, how the practice of doing research for development can be improved.

John Mayne is an independent advisor on public sector performance. Over the last 12 years he has focused on international development evaluation and results-based management (RBM). His current interests are on approaches for strengthening impact evaluation and useful theories of change in complex settings.

Cynthia McDougall is the Gender Leader at WorldFish and of the CGIAR Research Program on Fish Agri-food Systems (`FISH'). She is an interdisciplinary social scientist with over 20 years of experience in food security, natural resource governance, gender and social equity.

Rodrigo Paz-Ybarnegaray is an agricultural engineer with 20 years' experience designing, implementing and strengthening planning, monitoring, evaluation and learning systems particularly in America and Asia, but also in the Pacific and Africa.