Adaptive Evaluation: A Complexity-Based Approach to Systematic Learning for Innovation and Scaling in Development

Adaptive Evaluation:

A Complexity-Based Approach to Systematic Learning for Innovation and Scaling in Development

Siddhant Gokhale and Michael Walton

CID Faculty Working Paper No. 428
March 2023

Working Papers
Center for International Development
at Harvard University

IMAGO
Global Grassroots

Adaptive Evaluation: A Complexity-Based Approach to Systematic Learning for Innovation and Scaling in Development

Siddhant Gokhale and Michael Walton

Sep 9, 2022

Acknowledgements

We would like to thank several people for their input and continued support.

First, we would like to thank Isabel Guerrero from Imago Global Grassroots, who actively supported and encouraged us throughout—but especially in the early stages. This paper would not be possible without her involvement.

Second, we would like to thank the Harvard Kennedy School of Government faculty for their guidance and direction at important junctures. We would like to thank Rema Hanna for her precise suggestions for improvement, and questions which pushed us to search for answers. We would like to thank Janina Matuszeski who pointed us to academic papers which were crucial in placing an Adaptive Evaluation in the broader literature. Last but not least, we would like to thank Julie Wilson, Dan Levy, and Teddy Svoronos, who shared their insights and experiences in research design and methods and reviewed early versions.

Third, we would like to thank the experts whose advice we consulted. We would like to thank Monica Biradavolu (QualAnalytics) for her support in enriching our understanding of qualitative tools. Monica is an active advisor on two of Imago Global Grassroots' ongoing Adaptive Evaluation projects in India, and her contributions to those projects helped inform this paper. We would like to thank Vijayendra Rao (Research Development Group, World Bank), whose critical inputs helped us sharpen the language and framing of an Adaptive Evaluation, particularly in the context of change processes. From the Center of Policy Research in India, we would like to thank Yamini Aiyar and Mekhala Krishnamurthy, who generously shared their views on the early versions. We would also like to thank Arundhita Bhanjdeo, Nivedita Narain, and Surili Seth, coauthors of a qualitative study in Madhya Pradesh, which helped inform aspects of thinking in this paper. Last but not least, we would like to thank Adil Ababou, (Bill and Melinda Gates Foundation), who shared his experiences in evaluation and iterative implementation.

It would be remiss of us not to express our deep gratitude to the funders, partners, and project teams of the various Adaptive Evaluation projects and workshops Imago Global Grassroots has been involved with in India, Brazil, and Chile. These funders and partners have been invaluable in helping us learn about the best ways to implement an Adaptive Evaluation. The funders and partners include the Bill and Melinda Gates Foundation, the Self-Employed Women's Association (SEWA), Transform Rural India Foundation, Project Concern International- India, Sambodhi, IDinsight (especially Divya Nair, Sumedha Jalote, Neha Verma, and Christy Lazicky), Instituto Gesto, Fundação CAEd at UFJF, the Municipality of Recife, Brazil, and Fundación Colunga in Chile. Adaptive evaluation projects and workshops are an important laboratory for testing, consolidating, and refining our thinking. The learnings and lessons from these have undoubtedly shaped this paper. Imago Global Grassroots is currently involved in four Adaptive Evaluations— on the gains and viability of convergence of social and economic programs for the empowerment of rural women in two states in India, on the design and functioning of an enterprise support system to support the women-owned enterprises of SEWA in India, and two separate projects of varying capacities on facilitating back-to-school pedagogies in response to learning deficits due to the COVID-19 pandemic in Brazil. Special thanks to Siddharth Yadav, Ria Dutta, Priyal Patil, Bhavika Kohli, Gayatri Rao, Jossie Fahsbender, and Nicole Carpentier from Imago Global Grassroots for all their contributions to our Adaptive Evaluation projects.

Any errors or omissions are solely our own.

Abstract

Nearly all challenges in international development tend to be complex because they depend on constantly evolving human behaviour, systems, and contexts, involving multiple actors, entities, and processes. As a result, both the discovery and scaling of innovations to address challenges in development often involve changes in system behaviour or even system-level transformation. This is rarely a linear process over time and can result in unexpected outcomes. Existing evaluation techniques commonly used in international development, including Randomized Control Trials (RCT) and quasi-experimental methods, are good at assessing specific effects of interventions but are not designed for the change processes inherent to innovation and scaling within a system. There is a need to reconstruct how we use existing measurement tools, techniques, and methodologies so that they capture the complexity of the environment in which an intervention or change occurs. We introduce Adaptive Evaluation, designed to learn at various levels of complexity while supporting the transformation needed to foster sustainable change. An Adaptive Evaluation uses three main approaches to work with complex questions—systems diagnosis, theory-based assessment of change processes, and iterative designs. An Adaptive Evaluation typically builds hypotheses from field-based interactions, emphasizes learning over testing, advocates open-mindedness with techniques, and appreciates the value of dialogue and participation in navigating complex processes. It can use RCT or similar techniques to analyse specific processes within a system or a development cycle, but these are embedded in a broader approach to assessment and interpretation. It is designed to be flexible and adjust to shifting contexts. Finally, an Adaptive Evaluation can be applied at any stage in a complex intervention's lifecycle, from the interpretation of the system and change processes to rapid experimentation, prototyping, and testing of select interventions, and then adaptation to different settings for impact at scale. This paper provides the theoretical basis for an Adaptive Evaluation—the main approaches, core ideology, process, and applications.

Abbreviations

AE – Adaptive Evaluation

BPT – Bayesian Process Tracing

CLF – Cluster Level Federation

CMO – Context-Mechanism-Outcome

GP – Gram Panchayat

MoRD – Ministry of Rural Development

NRLM – National Rural Livelihood Mission

NREGA – National Rural Employment Guarantee Act

RDD – Rural Development Department

RCT – Randomized Control Trial

PRI – Panchayati Raj Institutions

PDIA – Problem-Driven Iterative Adaptation

SHG – Self-Help Groups

SEWA – Self-Employed Women's Association

VO – Village Organization

QCA – Qualitative Comparison Analysis

Figures

Table of Contents

1. Introduction

> "The method (of the social sciences) is … to associate the explanation of what exists with the imagination of transformative opportunity. Not some horizon of the ultimate possible but the real possible which is always the adjacent possible…And then, the vocation of social science is to help us understand *how* we came to be in this present situation, in such a fashion that our understanding of our circumstance…awakens us to the imagination of the adjacent possible."

(Unger, 2014)

Development is more than just the realization of positive outcomes for humanity (for example, increases in income, capabilities, and freedoms of people)—it is the process of change or transformation that induces these desired outcomes. Even the etymology of the word 'development' refers to a "*gradual unfolding*" or "*advancement through the progressive stage*", both of which are processes of change rather than a statement of mere improvement. Understanding *how* change occurs, more specifically, *how* a purposive action (often in the form of a 'policy' or an 'intervention') affects the desired outcome is the real prize of an evaluation for development understanding and action. By contrast, existing evaluations that have gained prominence in the last few decades, namely Randomized Control Trials (RCTs) and other quasi-experimental techniques (we refer to them as **standard impact evaluations** henceforth), primarily seek to understand whether a policy or intervention results in a realization of a positive outcome (albeit with robust causal attribution). While these impact evaluations have made important contributions, they are not designed to address the *central* question of development, which concerns the mechanisms and processes at work that result in the desired outcome (we refer to these as **change processes**).

Change processes are vital in development. Only if we understand *how* things become as they are (that is, how the state, society, and system functioning led us to our current state) would we be able to truly inform our actions about how to alleviate deprivation and poverty, enhance the individual and collective sense of self of people, and expand fundamental freedoms. Understanding change processes, thus, naturally involves delving into, engaging with, and navigating the highly **complex** environment that an intervention or policy inhabits, comprising of constantly evolving and unpredictable human behaviour, systems, and contexts. We suggest that there are two types of change processes that are of particular interest for development action—**innovation**, the process of designing, improving, and successfully adapting a novel idea in a local setting, and **scaling**, the process in which innovations are expanded to reach a substantial portion of intended beneficiaries. Facilitating innovation and scaling change processes in a complex environment that is constantly shifting, requires **learning** of a specific kind. This will sometimes involve interpreting large changes, but will often involve trying several variants of an intervention in small, frequently recurring steps, and then doing more of what works. The latter process mirrors the evolutionary process in biology, which through repeated cycles of genetic variation and natural selection helps species adapt to the changing external environment, while periodically leading to large shifts (Harford, 2011).

This paper introduces an Adaptive Evaluation—a complexity-based approach to systematic **learning** that is designed to support **innovation** and **scaling change processes** in development. An Adaptive Evaluation is an attempt to reconstruct how we use existing measurement tools, techniques, and methodologies in social science to (1) understand the complexity of the environment that an intervention or policy inhabits, and (2) support and inform action (i.e., policy and interventions) that induces purposive and meaningful change. Note that understanding how change occurs may include testing whether a specific intervention has a causal impact on a target group (for which RCTs are very well-designed), but this is only one small element of how change occurs (which is via the intervention in isolation). A significant part of the story of how change occurs is how the intervention interacts with the confluence of forces in the complex environment it resides in. This requires other techniques and a novel overall approach. For this reason, while an Adaptive Evaluation may include RCTs and related techniques for specific situations, in general, it incorporates a variety of, at times eclectic, techniques aligned with the questions raised by these types of change processes.

An Adaptive Evaluation builds onto the remarkable advances in the past three decades in empirical work and development thinking. Empirical evidence was not always central to economic and social science discourse. As recently as the 1980s, economist Edward Leamer observed how "*Hardly anyone takes data analysis seriously*" (Leamer, 1983, p.37). Since then, however, economics has seen a "credibility revolution" that is unprecedented, led largely due to improved quality of research designs, such as RCTs and quasi-experiments, along with the availability of more and better data (Angrist & Pischke, 2010). Perhaps most pertinently, putting hard numbers to key parameters of interest is now a common practice among policymakers, governments, donors, and social scientists. Similarly, it was not long ago that economists, policy advisers, and technocrats thought development challenges were simple—one just needs to get incentives right and follow 'best practices.' The Washington Consensus, first articulated by John Williamson in 1990, codified this into ten recommendations that broadly advocated for countries to "stabilize, privatize, and liberalize" (Williamson, 1990). By the mid-2000s, economists learnt the hard way that development is much more complex—a realization that stemmed from the underwhelming performance of the Washington consensus reforms in Latin America and Sub-Saharan Africa (Stiglitz & Schoenfelder, 2003; Rodrik 2006). A 2005 World Bank report, titled *Economic Growth in the 1990s: Learning from a Decade of Reform,* is highly illustrative of this shift in thinking. Based on the premise that development is complex, it prescribed a distinctly different approach, one that emphasized *humility*, acknowledging that there is no blueprint of what works that can be easily adopted, and *experimentation*, to select from a diverse mix of modest policy reforms (Zagha & Nankani, 2005). Finally, in the 2010s, economists from the Santa Fe Institute, among others, began drawing strong parallels between economics and complexity theory (which we describe in Section 2).

Together both the empirical and complexity revolutions in development have brought to the fore the next challenge— How to remodel existing tools of measurement in the social sciences to answer some of the more complex questions about transformation, in particular, change processes, such as innovation and scaling that are so central to development change? An Adaptive Evaluation attempts to answer this very question. It is a product of the improved consciousness and understanding in development over the past decades and seeks to build onto this strand of progress.

Central to the conception of an Adaptive Evaluation is the argument that the development and the change process it encompasses, namely innovation and scaling, are examples of complex adaptive systems. Intuitively this is because they consist of several actors, entities, and processes with multiple lines of action and consequences depending on interactions amongst multiple actors. We draw heavily from complexity theory to define the features of a complex adaptive system—as they apply to natural sciences, ecological systems, human behaviour, and the social, economic, organizational, and human systems of society (Byrne, 2013; Kast & Rosenzweig, 1972; Turner & Baker, 2020; Von Bertalanffy, 1972). Broadly, complex adaptive systems are *open* (i.e., they cannot be isolated from influences outside their notional boundaries), *nested* (i.e., they include subsystems are that also complex), *non-linear* (i.e., a slight change in one part may have a large impact on the overall system; conversely a substantial change may have little impact overall), *dynamic* (i.e., they have a trajectory and change with time), and *unpredictable*. Crucially, complex adaptive systems are characterized by *self-organization*, in the sense that there is no external impulse or guiding authority that controls them, and *emergence*, which refers to the property that the system forms an identity of its own that is distinct and cannot be deduced from its constituent parts. Perhaps most pertinently, complex adaptive systems, as the name suggests, are *adaptive*, with the ability to learn from and respond to changes. Finally, complex adaptive systems, involve feedback loops, in which changes in an element of the system cause the element itself to be altered.

Conceptualizing a system in this way makes it clear that development is fundamentally about engaging with complex adaptive systems (Barder, 2012a; Barder, 2012b; Beinhocker 2006). This applies to the education system, the healthcare system, the industrial, service, and agriculture economy, and the government, let alone the broader macroeconomy, and the global world order. Multiple behavioural, social, cultural, historical, contextual, and economic forces influence nearly any development challenge. This is especially true of innovation (Turner & Baker, 2019) and scaling. Innovation involves the creative exploration of new techniques, designs, and processes, with intrinsic unknowns on interactions, that in development, involve bringing such new processes into a human, societal and organizational context. Similarly, scaling involves taking a change that "works" in one part of an organization, specific geography, or a subsystem, into the new territory, with associated multiple interactions of individual, group, and organizational actors. Both are complex adaptive processes.

The fact that development, scaling, and innovation are complex adaptive systems has important implications for evaluations in development (Byrne, 2013; Walton, 2014). First, evaluations in development cannot simply evaluate parts of a system (typically a specific intervention) to interpret the behaviour of the system as a whole. Any evaluation based on complexity theory would aim to both (1) understand a specific part of the system (say an intervention), which may be evaluated using standard impact evaluation, but also (2) understand how the part (the intervention of interest) interacts with other parts of the system. Both steps are important— they are individually necessary, but only jointly sufficient to understand change processes. While this may seem, on the surface, a mere extension of a standard impact evaluation, it is a radical insight, one that challenges the universally accepted adage that to solve a complex problem, one can break it down into several simple problems. Much of development economics is focused on building models to provide guides to think about complex realities. From a complexity standpoint, this is insufficient, especially if it isn't complemented by other approaches and a deeper introspection of what it means to break out of the implicit and explicit assumptions of these models.

Second, a complexity-based evaluation cannot ignore the role of history and context (Walton, 2014). Any understanding of change processes, that is, how change occurs, must involve a careful examination of the trajectory of societal and organizational forces pertaining to the development problem. We never truly start from a blank slate, and history has embedded in it, features of a system's functioning that shape current behaviour, with lessons for the design of development action in the present. Third, any complexity-based evaluation understands that causality is rarely linear and uni-directional (Byrne, 2013). Last but not the least, an implication of the complex nature of development on evaluations, is that the nature and timing of impacts may not be known (Walton, 2014). To illustrate, an intervention, such as women collectives' groups, may take months, years, or even generations for transformational impacts on outcomes such as women's agency and empowerment to surface. Moreover, it is unclear if the improved agency and empowerment will surface in form of greater household bargaining and decision-making, greater mobility, an improved political franchise, increased sexual and reproductive freedoms, or a combination of these and more. We will demonstrate in the subsequent sections, how an Adaptive Evaluation in its design and process, deliberately incorporates each of these implications.

Many of the considerations of complexity in evaluations outlined above are overlooked in standard impact evaluations. Nonetheless, quite a few people have conceptualized the features, philosophy, and management of complexity-based evaluations in development (Bamberger et al., 2015; Byrne, 2013; Forss et al., 2011; Stern, 2012; Zimmerman et al., 2012), including identifying specific techniques and methodologies to address complexity. These methods include case studies (Paparini et al, 2021; Widner et al., 2022), process tracing, outcome harvesting (Wilson-Grau & Britt, 2012), and qualitative comparative analysis (Blackman et al., 2013). These are better equipped to analyse complex systems and are increasingly gaining prominence in some parts of development discourse. Complete evaluations based on tenets of complexity theory are rare, but a few exist. One example is Pawson and Tilly's Realist Evaluation, a theory-based approach which aims to understand the generative mechanisms and motivations of actors that caused an intervention's outcomes, specifically keeping in mind the specific socio-historical context (Pawson & Tilley, 1997). Another notable example is Patton's Developmental Evaluation, which is aimed at supporting innovation in dynamic and complex environments with emergent properties (Patton, 2010).

An Adaptive Evaluation's contribution to existing literature and thinking on complexity-based evaluations is twofold. First, an Adaptive Evaluation is more holistic and ambitious in how it incorporates complexity thinking. Existing evaluations based on complexity tend to emphasize specific aspects of complexity. In the case of a realist evaluation, this is the role of context and motivations in the realization of an outcome, while in the case of a developmental evaluation it is the role of innovation. An Adaptive Evaluation, in contrast, is more comprehensive in its scope: it incorporates, (1) an evaluation framework to understand histories, systems, contexts and mechanisms (similar to the goals of a realist evaluation, although different in methodology), (2) an evaluation framework for innovation, that involves iterative learning using rapid trial and error (that may feature in a development evaluation) and (3) a novel evaluation approach to scaling. Second, an Adaptive Evaluation offers a clear strategy, not only about techniques and methods to use in each type of framework, but under what conditions, and in what sequence is it most useful to use these tools. While most complexity-based evaluations, unlike standard impact evaluations, emphasize methodological flexibility and sometimes suggest techniques and methods that are useful, there is little guidance or thinking about when and how to use the variety of tools. In contrast, an Adaptive Evaluation offers three kinds of decision maps on methods, based on the type of question, the level of complexity and the stage in scaling and innovation. In this sense, an Adaptive Evaluation is more practical and implementable. Note that while the techniques and methods highlighted in an Adaptive Evaluation are not in themselves novel (they are adopted from design thinking, the social sciences, and economics), the combination of techniques to use, and their sequence is a contribution.

There are three main methodological elements of an Adaptive Evaluation elaborated in this paper: diagnosis of system functioning; theory-based exploration of change processes; and iterative feedback to intervention design. We briefly introduce these here and also highlight a few of the specific methodologies that are particularly useful. Detailed accounts of both concepts and methodologies are in the main text.

Systems diagnosis involves the interpretation of how a system functions. This involves understanding both the relationships within a system and overall system behaviour. A core practical technique we have developed in fieldwork is the building of a system map through the participatory engagement of actual actors, utilizing design thinking techniques. This can, for example, be built from a listing of actors, followed by asking a carefully curated group of stakeholders the question of how each actor relates to others, as well as how they see the power, incentives, and culture of different elements of the system. Then, as a crucial step in understanding potential change processes, this moves to an assessment of sources of leverage or blockage within the system. A system mapping workshop of this kind can fruitfully be integrated with histories, literature, and academic work on the system. For example, in a system diagnosis involving the state, this would involve academic and other interpretations of bureaucratic behaviour and interactions between the state and other social actors. There are varieties of additional specific techniques that be used in systems diagnosis, notably network analysis and mathematical systems dynamics of varying degrees of sophistication. Formal modelling has the advantage of clarity for the modeller, but the disadvantage is a loss in flexibility and often greater opaqueness for other participants. While a system diagnosis is ideally a first step and initial foundation of an Adaptive Evaluation, it would typically be explored, fleshed out, and updated, as one develops a greater understanding of the system, and indeed, as the system itself changes.

Theory-based assessment of change processes is then a critical element in the evaluation of interventions within a system. It typically engages with only selected pathways of change within the system and starts from building what may look like a conventional theory of change, supported by an (often linear) logic model structure. Then a key element is theorizing the steps within such a model in ways that can articulate testable hypotheses that can be taken to available data, allowing for updating of the specific pathway and other influences of consequences. While this indeed involves hypothesis testing, this is in the spirit of dynamic learning. Where feasible and practical this can certainly use standard statistical techniques, including randomization of specific treatments. However, often this will not be feasible, and techniques of process tracing, using logical tests of specific theory-based hypotheses of causal relationships, are more appropriate. These are more amenable not only to relatively frequent updating of interpretations but also to inference from a broader variety of sources of evidence.

Iterative processes involve integrating the discoveries and interpretations from the evaluation into the intervention design itself. The idea is to engage in a cycle of continuous experimentation, testing, learning and refinement. Naturally, this involves close coordination with actors on the intervention side. It requires alignment with the mindset of such actors—that is itself part of the system diagnosis—and with the expected time dimension of change. Some iterative processes can be quite rapid, where responses are expected and can be tracked over weeks and months. In these cases, an Adaptive Evaluation can be tightly linked with agile processes on the intervention side, bringing the theory, hypotheses testing and measurement approach of an evaluation into design sprints or other adaptive processes on the implementation side. Process tracing, A/B testing of alternative designs, or the more in-depth interpretation of relative successes ("positive deviance" analysis) are examples of techniques that can be applied to relative rapid iterative design processes. Where the counterparts for implementation have different organizational cultures and mindsets, as in many governmental systems, the Adaptive Evaluation can be "adapted" to the incentives and processes within these.

While these three elements have strong complementarities, the actual mix, and especially the specific methodologies will depend on the nature of the problem. We discuss in the text an "ideal" or archetypal sequence in an innovation and scaling process, but actual experiences will typically be messier. Theories of change have to be updated in the transition from a very local, perhaps tightly supervised or highly motivated subsystem, to the broader system.

While Adaptive Evaluation proposes a very practical set of techniques to address complex questions, this is embedded in values and an underlying philosophy about development and evaluation—one that embraces complexity and contradictions, focuses on change processes and not just outcomes, brings people together to co-create solutions, and is based on empowering people closest to the problem to take ownership of their situation. It is not only what we do in an Adaptive Evaluation in terms of methods and techniques, but how, and with what underlying values the methods and techniques are used, that matters immensely.

Thus, an Adaptive Evaluation does not presuppose a hypothesis and fix outcomes of interest ex-ante but investigates multiple hypotheses on processes and outcomes that emerge from work in the field. This ensures that evaluators are open to adapting hypotheses to emerging realities of the field, rather than imposing a fixed hypothesis. Shifting hypotheses can make an evaluation less amenable to statistical causal testing in one period, but it makes it more useful for the examination of change processes. Adaptive evaluations are participatory with an emphasis on co-creation. This is because actors are seen as key sources of knowledge. An Adaptive Evaluation understands that beneficiaries closest to the problem understand their circumstances better than any evaluator or expert. Moreover, any understanding of change processes requires participants of the change process to be actively involved to take ownership and drive the change, but also to use the evaluation to inform their actions. Finally, in contrast to standard impact evaluations, an Adaptive Evaluation emphasizes learning over testing and uses a variety of techniques.

We use 'adaptive' very intentionally. First, 'adaptive' highlights the recognition that development, scaling, and innovation all share features of complex adaptive systems. Second, 'adaptive' refers to the emphasis on continuous learning and improvement to adapt intervention design to the evolving needs of the end beneficiaries, and to adapt to new contexts and settings as the intervention scales. Third, the 'adaptive' points to the fact that evaluation itself is adaptive, in that the methods and techniques used are intended to adapt to the nature of the change process (innovation or scaling), and the complexity of the system.

Last but not the least, 'adaptive' points to two other elements of the change processes that are outside the scope of this paper but are nonetheless crucial complements to an Adaptive Evaluation—adaptive leadership and adaptive implementation. Adaptive leadership creates the enabling environment for meaningful change to take place(Heifetz et al., 2009). Adaptive implementation involves putting in place the structures and mechanisms for rapid prototyping, experimentation, and learning. This is at the centre of agile processes, and sequential design sprints in the private and public sectors; while a closely related approach for the public sector is the Problem Driven Iterative Adaptation (Andrews et al, 2017). While we acknowledge these as complementary adaptive processes, we do not delve into them in the rest of the paper.

It is important to clarify that we do not use 'adaptive' in the very specific lens it is most recently used in the econometrics literature, in the (very interesting) approach to adaptive experimental designs. Adaptive experiments involve testing multiple treatment arms over a series of waves. Each subsequent wave reorganizes the allocation of participants across the multiple treatment arms based on learnings from prior results in a way that prioritizes the most effective treatments. Unlike an RCT, which answers "does the intervention have a significant effect?", an adaptive experiment answers "which intervention has the greatest effect?" The econometrics research on adaptive experimental designs focuses on improved algorithms for the reallocation of participants (Kasy & Sautmann, 2021) and better methods for statistical inference (Hadad et al., 2021). While adaptive experiments are very much in the purview of the spirit of an Adaptive Evaluation (and is certainly a technique that we can use), our use of 'adaptive' is much more general, and not confined to experimental methods to update RCTs.

The rest of the paper is organised as follows. **Section 2: Complexity, Development, and Evaluation** provides an in-depth understanding of complexity theory, its application in development, innovation, and scaling, and its implications for evaluations. **Section 3: Adaptive Evaluation**, develops a framework for evaluation questions based on the Cynefin Framework, which is based on complexity theory, and organizes problems into those that are "simple", "complicated" and "complex"; this formally introduces an Adaptive Evaluation as appropriate to complex problems, and then describes the methodological elements of an Adaptive Evaluation. **Section 4: The Adaptive Evaluation Process: From Innovation to Scaling**, provides a practical guide to the Adaptive Evaluation process, summarizes methodologies and links these to the conceptualization of innovation and scaling. It also suggests an ideal sequence or process for an Adaptive Evaluation and develops the idea of an evaluation as a journey. Throughout, the paper we use numerous real-life examples and thought experiments, intended to be accessible to any reader engaged or interested in development.

2. Complexity, Development, and Evaluation

> "Science has explored the microcosms and the macrocosms; we have a good sense of the lay of the land. The great unexplored frontier is complexity"

(Pagels, 1989, p.12)

The objective of this section is to lay the foundations of the underlying theory that informs the elements, process, and applications of an Adaptive Evaluation. An Adaptive Evaluation draws from complexity theory and uses a complexity frame of reference for development. This section defines and introduces key concepts and features of complexity theory, the links between complexity theory and development, and the implications of a complexity frame of reference on evaluations in development.

This section has three sub-sections. Section 2.1 defines complexity theory as the study of complex systems, describes features of complex systems, and introduces the characteristics of a special class of complex systems—complex adaptive systems. Section 2.2 makes the argument that nearly every problem in development resides in a complex adaptive system. Further, it illustrates how innovation and scaling can also be conceptualized as a complex adaptive system. Section 2.3 looks at the implications of complexity for evaluations in development.

2.1. Complexity Theory

2.1.1. Systems and Complexity

Complexity theory studies the characteristics and features of complex systems (Strathern & McGlade, 2014).

A **system** is a set of actors, entities, or processes that form a unified whole through their various configurations, dependencies, relationships, and interactions (Turner & Baker, 2019). Any system, by construction, has a boundary (Kast & Rosenzweig, 1972). Within its boundary could be more embedded systems, and outside the boundary, the system itself could be contained within other larger, encompassing systems (Byrne, 2013). For illustration, consider a public school to be a system. Within the school are systems — different learning levels organized by grades (primary, middle, and secondary), different departments organized by subject (math, science, language, humanities, arts) and different offices organized by function (admin, finance, HR). The school is part of a larger system of education, with a curriculum from a local or national board, a community of parents, a supply of teachers, etc. Moreover, the school itself resides in a larger environment and context, say, in a rural area with high illiteracy.

But complexity theory concerns itself mostly with **complex systems**. A complex system is composed of several actors, entities, components, and processes, each with (1) multiple lines of action, and (2) consequences dependent on interactions between multiple actors (Gare, 2000). Complex systems are difficult to model because they consist of conflicting incentives, imminent tensions, intricate relationships, critical dependencies, powerful competitions, and vigorous interactions between actors, entities, and processes and between the system and its external environment (Morçöl, 2013).

In a complex system, the 'whole' is distinct from the properties of 'individual' elements. This is because system-level behaviour is also determined by the interactions *between* actors rather than the features of individual actors themselves. For example, a government is a complex system. Governments viewed as organizational entities have certain characteristics: they are rigid, bureaucratic, hierarchical, and susceptible to leakages, delays, and corruption. A government system's behaviours and traits are not determined by the sum of behaviours and traits of its departments, front-line service delivery agents, technocrats, and ministers. Instead, the government systems' characteristics derive not only from their constituent parts but also from the sum of the resultant interactions and incentives between departments and actors (Von Bertalanffy, 1972). As a result, one cannot infer the behaviour of the government from the behaviour of its constituent parts.

Our world is replete with examples of complex systems. Throughout Section 2, we will focus on one very specific example of a complex system, the Rural Development Department (RDD), which exists within the state governments of India, and is linked to the Ministry of Rural Development (MoRD), a branch of the central government.[1] We focus principally on two flagship programs of MoRD, the National Rural Livelihoods Mission (NRLM) and the Mahatma Gandhi National Rural Employment Guarantee Act (NREGA). Both NRLM and NREGA aim to increase the income of rural households through local community institutions, and both focus on women's participation and economic empowerment. The RDD implements and executes the mandate of the MoRD, and its flagship programs, in the state governments of India.

NRLM, at its core, improves women's access to finance through federated structures of women's Self-help Groups (SHGs), that essentially pool small funds to reduce risk as in a classic micro-finance model. Additional activities, such as health and nutrition interventions, livelihood skills training,[2] and support for market access, are often layered onto the SHG system as they mature which further relates to the goal of women's empowerment. NREGA guarantees each rural household 100 days of paid employment per year at a pre-specified wage. The paid employment in NREGA is usually manual labour to create productive, income-generating assets and infrastructure for village development (such as dams, wells, irrigation systems, roads, cowsheds, schools, nurseries etc.). NREGA is often used to provide income outside of the harvest seasons when there are few employment opportunities.

Many RDDs in the state governments are attempting to link the largely independent efforts of NRLM and NREGA, a process known in policy circles in India as convergence. Convergence is motivated by some natural complementarities between NRLM and NREGA. For example, increased income from NREGA potentially gives rural households more confidence to repay loans, and thus increases the demand for credit from SHG's part of NRLM. On the other hand, NRLM reduces the cost of borrowing for women, which creates a demand for NREGA wage funds to help repay loans. In addition, productive assets and village infrastructure created from NREGA, can boost local productivity, and increase incomes, which further encourages borrowing from SHGs via NRLM. In this way, NRLM and NREGA are mutually reinforcing mechanisms that could be potentially leveraged to alleviate rural poverty. State government RDDs are currently experimenting with interventions to foster convergence between the two programs.

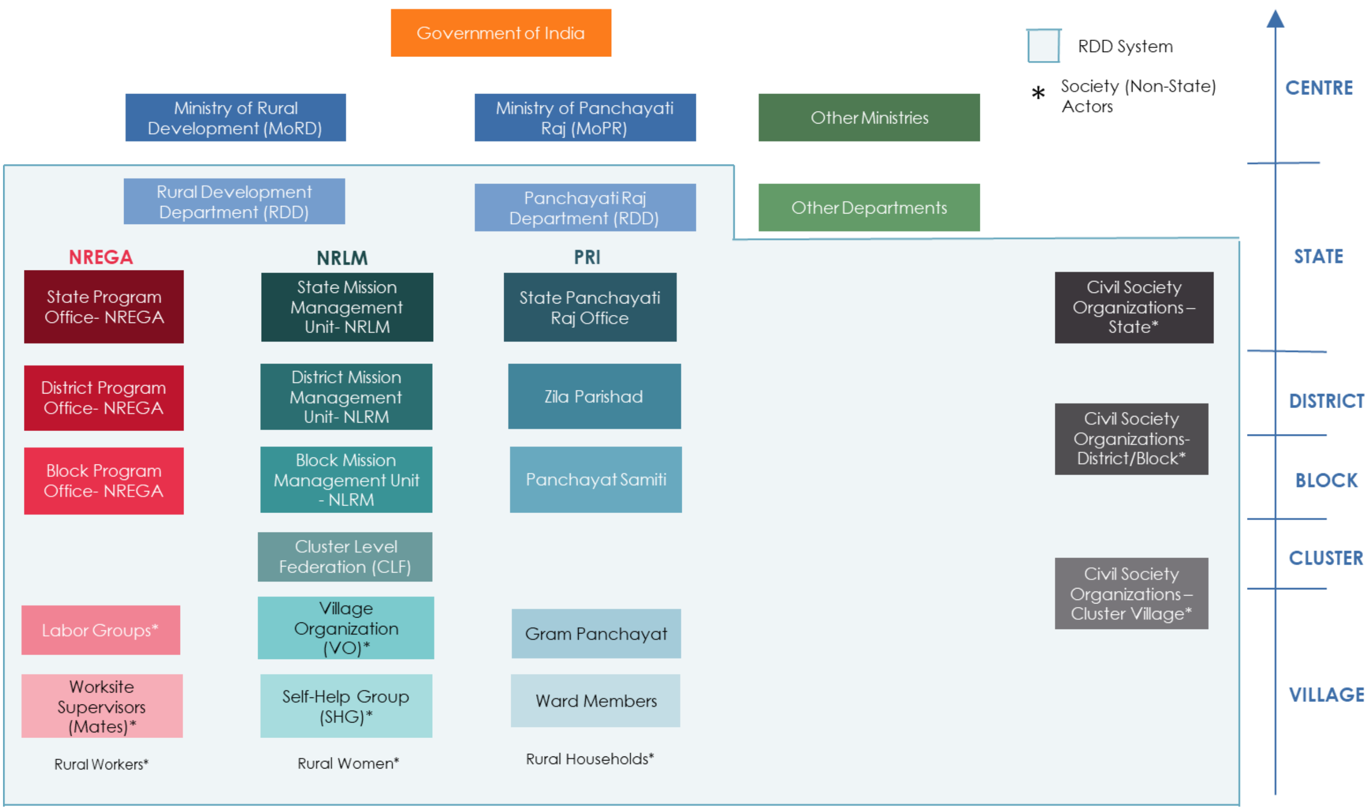
The RDD can be viewed from the lens of a system. For sake of clarity, let us define the boundaries for what we refer to henceforth as the **RDD System**, which is much broader than simply the RDD in the state. Within the RDD System in a state are sub-systems (or departments) responsible for NREGA, NRLM and other national schemes, but also the Panchayati Raj Institutions (PRI), through which the decentralized and participatory self-government of the villages in rural India is realized, under a parallel department to rural development. However, in addition to sub-departments overseeing NRLM, NREGA, and PRI, the RDD System also contains within its societal actors and entities, such as workers, labour unions, SHGs, and civil society organizations. Outside of the RDD System are other departments of the state (such as health, finance, social welfare etc.), the larger state government, the MoRD at the centre, and the larger central government. See **Figure 1** for an illustration of the RDD System.

[1] We use the term 'central government' or 'centre' to refer to the federal or national government (i.e., the Government of India). This is common parlance in India.

[2] e.g., training for sewing school uniforms, training for sewing masks, training to produce hand soaps and sanitisers, training to run community kitchens, etc.

Needless to say, the RDD System is complex. It has several actors, entities and processes that eventually lead to service provision for the poor. The most basic entity in the NRLM system is an SHG, which typically consists of 10 to 20 rural women, who together can borrow via formal banking institutions. 10 to 15 SHGs form a Village Organization (VO) typically responsible for developing investment plans. VOs often serve as the main channel for providing Community Investment Funds (CIF) and link SHGs to the banks. 25 VOs typically form a cluster level federation (CLF), which liaise and coordinate with block government officials of NRLM, and the elected official (called a Pradhan) of the village governing council, the Gram Panchayat (GP). The Gram Panchayat is one part of a three-tier Panchayati Raj Institution (PRI), which also includes the Panchayat Samiti (at the block level), and the Zila Parishad (at the district level), which oversee and manage programs, implementation, and monitoring of the Gram Panchayat. The Gram Panchayat, responsible for the implementation of central and state government schemes, economic development, and the strengthening of social justice in the village, coordinates, among many things, the creation of job cards for NREGA eligibility, and the planning of projects for NREGA work activities. The NREGA block and district officials coordinate with the Gram Panchayat for the implementation of NREGA. Finally, directing the individual block and district NRLM and NREGA program managers, are state heads of NRLM and NREGA, all of which are part of the RDD, and responsible for policy decisions and guidelines.

Figure 1: The RDD System



The different parts of the RDD System —NRLM, NREGA, and the PRI, interact in different and complex ways. For example, an SHG woman might be a leader in a VO, and also a mate (worksite supervisor) for an NREGA project to build a village road, that hires 40 workers paid Rs 192 per day for their labour. This SHG woman might go to the Pradhan of the village for any grievances that happen at the village level (e.g., to inquire about a delay in NREGA wage payments or report discrepancies in the muster roll). The Chief Minister of the state may pass a guideline that calls for 100 per cent women mates at NREGA worksites. The Pradhan may be sceptical of the guideline as it relinquishes his/her control of the worksite but might be forced to comply because a senior government official came for an unexpected inspection and reprimanded the GP for disregarding the policy guidelines. There are so many interactions, tensions, relationships, competitions, and dependencies in just this one snapshot.

In the following section, we will discuss eight main features of complex systems, using the RDD System in state governments in India as a detailed recurrent example to explain each feature. Nonetheless, we will also describe how each feature applies to diverse disciplines, from mathematics to the natural sciences, and the social sciences. Complexity theory truly transcends fields and is generalizable to many contexts.

2.1.2. Features of Complex Systems

Complex systems exhibit openness, nestedness, non-linearity, dynamism, unpredictability, cyclicality, self-organization, and emergence (Turner & Baker, 2019). While these are presented distinctly, it is important to recognize that, in reality, they are interconnected and reinforce one another.

I. Openness

Complex systems are open. Open systems cannot be truly isolated from their boundary because they are influenced not only by internal forces within the system but also by external forces of the overarching environment (Room, 2011).

In the natural sciences, an open system refers to a system that has a free exchange of energy, matter, and resources. In physics, an example of an open system is our solar system, which is influenced by forces within it, but also by a larger galaxy of stars, interstellar gas, dust, and dark matter. In biology, an example of an open system is an ecosystem, which consists of the organisms that inhabit it, but also the external physical environment they interact with. In the social sciences, another example of an open system is a city. Cities exhibit openness because they are influenced by internal forces (crime, housing, crowding etc.) but also external forces (migration, global warming, fiscal issues, pandemics etc.)

The RDD System is open. It is influenced by interactions within the RDD System, such as those between its key actors, (e.g. SHG women, workers, and government officials), entities (e.g., the NRLM, NREGA, and PRI departments), and processes (i.e., the rules, guidelines, and procedures). But it is also substantially influenced by interactions with the environment outside of the RDD. The RDD is influenced by other related departments, such as Agriculture and Famer's Welfare, Women and Child Development, and Finance, due to its intersectionality with different areas. The RDD System is also affected by global economic headwinds and tailwinds (e.g., commodity prices, inflation, and supply chain disruptions, among others). Anything related to rural development is inextricably linked with the weather, and in India, the monsoons, have a considerable bearing on the abundance of the harvest (Gokhale, 2020). Finally, all departments and how they operate cannot be removed from the politics and ideology of the majority in the national and states assemblies. All these external forces have an impact, and thus any understanding of the RDD cannot fully exclude these varied external forces.

II. Nestedness

A complex system is nested, that is, the elements or subsystems within the system are themselves complex systems (Byrne, 2013).

Nested systems are everywhere. The human body is a nested system, comprising complex subsystems, such as the circulatory system, digestive system, immune system, respiratory system, and nervous system. In the social sciences, economies are complex systems, because they consist of sectors (e.g., finance) and industries within it (e.g., commercial banks, investment banks, asset management companies, private equity, and brokerages), each of which is a complex system.

The RDD System exhibits nestedness. Every subsystem within the RDD System is a complex system, from the NRLM and NREGA departments to the PRI, and finally the villages and federated structures of self-help groups.

III. Non-Linearity

A complex system is non-linear, meaning a slight change in one part of a system can have a large impact on the overall system. Conversely, a substantial change in one part of a system could have a negligible effect on the system as a whole. More precisely, the size of the change in one part of the system is not proportional to the change in the overall system (Mitchell, 2009).

Non-linearity is an important concept. In mathematics, non-linear systems, are a set of simultaneous equations in which a change in the size of the input does not correspond to a proportional change in the size of the output (Strogatz, 2018). Non-linear systems also inhabit our natural world. In fluid dynamics, the motion of viscous liquids is described by the Navier-Stokes equations, which are known to be non-linear (Strogatz, 2018). Predator-prey dynamics are often modelled in biology using the Lotka-Volterra equations, which is a non-linear system (Brockmann, 2018b). In the social sciences, traffic jams are highly non-linear (Brockmann, 2019). A small event, such as deer crossing the road, can interrupt smooth traffic, causing all vehicles to slow down and create a bottleneck. The slowdown in speeds gets amplified as more cars arrive.

The RDD System is non-linear. A small, seemingly inconsequential event, such as a slight change in the heat differential between the land and the oceans (due to the atmospheric changes caused by global warming), could delay the monsoons, exacerbating farmer anxieties, or worse, result in torrential floods (Gokhale, 2020; Schwartz, 2021). The damaged harvest could reduce farmer incomes by 20 to 25 per cent in unirrigated areas (Hari et al., 2018), which would lead to increased demands for NREGA, and potentially an increase in the NREGA budget. The slight change in the heat differential has grave and devastating consequences for the RDD System and its actors. Another example comes from the COVID-19 pandemic. A small event, such as an obscure coronavirus infecting one person, can disrupt the entire RDD System, causing millions of workers to return to their villages without any prospect of income, stalling local panchayat elections, and massively increasing demand for NREGA.

IV. Dynamism

Complex systems are also dynamic: they are a function of time, and constantly evolve in a non-linear manner (as described above). The trajectory of a dynamic system depends not only on initial starting conditions but also on the past (De Domenico &

Sayama, 2019). In other words, complex systems convey histories and are sensitive to the current context, but also the factors that led us to the current context (Turner & Baker, 2019).

Most systems are dynamic and change over time. In biology and the social sciences, population growth is a dynamic system. In physics/mathematics, dynamic systems refer to a change in a particle's/point's state over time, which can be described through differential equations. The motion of celestial bodies is one among many dynamic systems in physics.

The RDD System is dynamic. To illustrate let us unwind the clock and look at the evolution of one of the flagship programs under the MoRD—NRLM (Ratan, 2022). The roots of NRLM can be traced back to several self-organizing efforts, notably the formation of the Self-Employed Women's Association (SEWA) in 1972. SEWA's founder, Ela Bhatt organized poor women, typically from historically marginalized (Scheduled Caste and Scheduled Tribe) communities who worked in informal jobs (such as weavers, potters, hawkers, and artisans, among others), with the aim of improving their income opportunities. By the 1980s various other non-governmental and civil society organizations (e.g., PRADAN, MYRADA), enabled the formation of the first Self-Help Groups (SHGs), which were remarkably similar in formation to the SHGs under NRLM today (Ratan, 2022). In 1982, the National Bank for Agriculture and Rural Development (NABARD) ran a pilot program to link SHGs to rural banks. Today this is the world's largest microfinance project. In 1999, the rising prominence of SHGs caught the attention of the MoRD, which launched the Swarnajayanthi Gram Swarozgar Yojana (SGSY), aimed at enhancing self-employment and skill-development through SHGs, and thus going beyond simply financial linkage. By the early 2000s, the SHG linkage with banks, layered with skill development and training expanded in state governments with positive impacts, particularly in Kerala and Andhra Pradesh (Ratan, 2022). This momentum led to the central government launching NRLM in 2011, with the mandate to leverage SHGs to improve rural livelihoods across the nation, supporting state-level versions of the NRLM in state governments that did not already have equivalent structures.

This evolution illustrates how several social factors coalesced over time, building in an incremental, iterative manner, to form the NRLM as we know it today. The NRLM is the product of a dynamic system, which depends on the context and the forces of history.

V. Unpredictability

The non-linear and dynamic nature of complex systems results in a new feature, unpredictability.

The unpredictability arises from a high sensitivity to initial conditions: a phenomenon known as **chaos** in mathematics (Lorenz, 1972). Chaos causes tiny changes in initial conditions to lead to dramatically different outcomes. Chaos theory shows that even deterministic models (i.e., models with no randomness, where how initial conditions lead to an outcome is exactly specified) can give wildly divergent predictions (Gleick, 2011). Long-term predictions in a chaotic system are almost impossible to model numerically because small rounding errors at any stage in the computation can produce unintended and inaccurate results.

Simple, deterministic systems, like a pendulum with two hinges (a double pendulum), are chaotic, that is, they do not ever follow the same path of motion over time, despite starting from almost exactly the same position or initial condition (Brockmann, 2018a). This is because you can never truly start in exactly the same position (there is always some error, albeit, small or even microscopic). A single pendulum, with one hinge, is a linear dynamic system that has no chaos, so its path of motion can be easily predicted. But a double pendulum with two hinges becomes a chaotic non-linear system that is impossible to predict. This can be attributed to the interaction between the two pendulums in the double pendulum, which create this unpredictability.[3] If simple, deterministic systems with chaos are so radically unpredictable, one can only imagine the levels of unpredictability in more complex, non-deterministic systems. In many situations in real life, there is randomness, and we do not know how initial conditions produce outcomes.

Consider the RDD System. The RDD System depends on some set of initial conditions (the economy, the climate, the politics, to name a few) that determine its trajectory. However, we do not currently know how exactly the initial conditions map onto the size, formation, structure, and functioning of the RDD, that is, the RDD System is non-deterministic. And yet, we know qualitatively the initial conditions have a significant bearing on the RDD System. All these factors make predicting how the RDD System will evolve impossible. NRLM, for instance, as a program of the MoRD, intends to only continue implementation at the block level for 10 years (*Implementation of NRLM*, 2022), after which it hopes community federations (especially the Cluster Level Federations) will take over. Will NRLM cease to exist in 10 years? No one can truly predict this. This is just one illustration of how the RDD System, like many complex systems exhibits unpredictability.

There are several examples of unpredictability and chaos[4] in other disciplines (Biswas et al., 2018). The weather is a common example of chaos, and in fact, Edward Lorenz, the founder of modern chaos theory, was a meteorologist. Chaos theory explains why long-term weather forecasts are impossible. In biology, epileptic brain seizures and cardiac arrhythmias (irregular heartbeats) are examples of chaos. In finance, stock market volatility exhibits chaos. In general, chaos has come to represent the idea that tiny choices or changes in initial circumstances can have huge effects.[5]

VI. Cyclicality (Feedback loops)

Complex systems have feedback loops. Feedback loops are effects on elements of a system that cause the element itself to be altered.

[3] The famous three-body problem, in mechanics, follows a similar pattern: one can predict the earth's motion around the sun (i.e., over the two bodies), but adding another body (say, the moon) renders calculation impossible. Newton famously agonized over this problem, and only later it was confirmed there is no simple solution to this problem due to its chaotic nature. Solutions to the three-body problems were found only recently (Cartwright, 2013).

[4] Chaos is sometimes known as the **butterfly effect**. More formally the phase space (the path of a particle) of the Lorenz system of equations resembles a butterfly. Colloquially, this is the idea that a flap of a butterfly's wing in Brazil. can cause a tornado in Texas, which incidentally is the catchy title of the paper that introduced chaos theory (Lorenz, 1972). This idea has resonated and is often used in popular culture (in movie titles, fiction books, etc.)

[5] The 2016 biographical drama, Lion, shows the journey of a 5-year-old boy, who was separated from his poor family after getting lost on a train in a suburb of Khandwa, Madhya Pradesh. He arrived at Calcutta and was eventually adopted by an Australian family and grew up in Hobart, Tasmania. This led to vastly different life outcomes for him relative to his siblings, another illustration of the sensitivity of initial conditions.

Feedback loops are common in nature. An example is the melting of the arctic ice feedback loop. As temperature rises, the arctic sea melts, causing reflective ice to disappear, leaving the darker ocean to absorb more heat, and thus further increasing the temperature. Economics also has many feedback loops. An example of a negative feedback loop is the poverty trap. Being poor, means less healthcare, education, and bandwidth, which then, in turn, exacerbates the state of poverty.

The RDD System and the complex subsystems within it include feedback loops. One example comes from the functioning of an SHG. An SHG works on the principle of members making small contributions. Say, for sake of illustration, we start with 50% of women in an SHG contributing Rs 20 each. The contributions lead to capital which when aggregated can secure loans. With the loans, women can invest in productive assets. The rise in income helps the women repay the loans. This repayment helps them get even bigger loans, as they demonstrate the ability to repay. The women who did not contribute notice the benefits of the system, and now 70 per cent of women contribute, and they contribute Rs 40, as they have more income. This is an amplifying or positive feedback loop; the loop keeps increasing the portion of the SHG members contributing and the size of the contribution, which increases the size of loans.

Now consider another feedback loop. All SHG women contribute Rs 100. The contributions lead to capital to help acquire a loan. The loans help women invest in productive assets. However, at the time of repayment, some members are unable or refuse, to repay. Other members do not repay in retaliation for the lack of payment by these members. The lack of repayment leads to the bank account of the SHG women being frozen. Fewer women now contribute, and they contribute less. The lack of repayment means they cannot get further loans. This is a dampening feedback loop: the loop keeps decreasing the portion of the SHG members contributing and the size of the contribution, which decreases the size of loans. This could cause the SHG to no longer function.

The rise and fall of SHGs in NRLM depend on these amplifying and dampening feedback loops, which could be triggered by forces of cooperation or non-compliance.

VII. Self-Organisation

Complex systems are self-organizing (Biswas et al., 2018). Self-organizing, in the specific sense of complexity theory, means that a system has its own internal dynamic of structure and change, a product of the multiple interactions within its whole, as opposed to being designed, shaped, and managed by any central actor or agency with absolute or direct control. This may lead to unpredictable system transitions, including in some cases striking manifestations of organization. This is sometimes called *spontaneous order* as it can organically lead to an emergence of organization in an initially disorderly system. Spontaneous order typically emerges because of a series of amplifying positive loops driven by highly localized actors that eventually influence the collective. Change that comes from self-organization is self-sufficient and more sustainable because it evolves organically and requires no external oversight to function. Decentralization and participatory bottom-up approaches provide an environment for self-organization.

Self-organization or spontaneous order is ubiquitous in the natural world. It can be seen in the radial symmetry of snowflakes, caused by local forces between water molecules and their environment (Ball, 2001). Ant civilizations (and in general herd behaviour[6]) are also examples of spontaneous order, in which remarkable coordination occurs in the absence of any centralized planning (Camazine et al. 2020). In physics, atmospheric circulation via convection is organized into smooth weather patterns. In social sciences and particular economics, spontaneous order can explain the formation of cities and business cycles (Krugman, 1996). Adam Smith's concept of an invisible hand also exhibits spontaneous order as people acting in their self-interest without central authority reach an equilibrium that benefits society (Krugman, 1996). This idea is also used to promote the idea that freedom can spawn technological progress (Tao, 2016).

The RDD System has self-organization. Ironically, the view from the core government, at central or state levels, is often of a top-down managed system, exemplified by multiple layers of rules, regulations, and targets. In actuality, this is far from the case: neither within the purely governmental parts of the system (the NRLM and NREGA departments in a State such as MP or Bihar) nor, even less, in the interactions with the societal elements of the system, do any single actors, including higher levels of government, manage the organization as ostensibly designed (Aiyar & Bhattacharya, 2016; Bhanjdeo et al., 2021; Pritchett, 2009).

As noted above, from a longer-term perspective, we can also see the current RDD System as the manifestation of a set of local organic processes initiated several decades ago that seemed disparate and disconnected, but slowly combined to create formal institutions mandated by the central government. In the nascent years of India's independence, poor rural women were scattered across the country, with little voice or work and no instrument to leverage their collective concerns. Then, social mobilizations of small groups of women (notably in SEWA and MYRADA) spread to activist civil society organizations inducing and nurturing women's groups (for example in the work of PRADAN). Eventually, state-level leadership saw an opportunity in state-organized women's groups, for developmental and political ends (as in the Andhra Pradesh Indhira Kraanthi Patham or District Poverty Initiative Programme). This was then replicated at a national scale with organizational and political guidance from the central government (with the formation of NRLM)—the last two phases in part supported by the World Bank.[7] Many actors were intentionally making decisions along this path, but the overall trajectory was one of a dynamic system that was *self-organizing* in the sense of complexity theory.

VIII. Emergence

Emergence is the idea that a system develops its own behaviours and patterns, which cannot be deduced from its parts (Goldstein, 1999). All the features above, openness, nestedness, non-linearity, dynamism, cyclicality, and self-organization contribute to the emergent properties of a system

Emergence has many examples in the natural sciences. A living organism has properties distinct from the multiple cells it is made of. Similarly, the brain has different properties from the billion neurons within it that help produce consciousness. From the Big Bang to the folding of proteins, everything is emergent (Pines, 2014). In the social

[6] Other examples include birds flocking, the synchronization of fireflies, and fish schooling

[7] The World Bank is itself an organization that incorporates an irreconcilable tension between top-down, engineering development thinking with aspirations for participatory societal engagement.

sciences, an example of a system with emergence is the internet. Social networks (like Twitter, Facebook, etc) often have their characteristics (more extreme emotions, exaggerated life outcomes, a tendency for positive self-promotion, etc.) that are quite different from the individuals that are part of it.

The RDD System is emergent. For example, SHGs in their structure are meant to be inclusive, participatory, and democratic. However, while the SHG is the basic unit of NLRM, the same cannot be said of the NRLM department, which is much more hierarchical. Parts of a system cannot describe the whole.

2.1.3. Complex Adaptive Systems

A complex adaptive system, in the literature of complexity theory, is a special class of complex systems with a key distinguishing feature: the capacity for adaptation (Holland, 1992b). Biology and the social sciences have innumerable examples of complex adaptive systems.

I. *Adaptation*

Adaptation, in complexity theory, is the ability of a system, and its constituent actors and entities, to respond to changes and learn from experiences, interactions, patterns and dependencies within and outside the system (De Domenico & Sayama, 2019). A consequence of adaptation is a system that is constantly evolving and transforming into new formations in response to previous formations.

Nature is replete with examples of adaption (Holland, 1992a). Our immune system is a pertinent example: it continually responds to and learns about pathogens to trigger a defence response. Similarly, our brain continually responds and learns from our external environment. In biology, adaptation refers to an evolutionary process, that via natural selection, enables an organism to be more able to survive in its habitat (Dobzhansky, 1968). Adaptation can explain the diversity of species we see in the world today. In computer science, artificial intelligence, and machine learning, are intended to develop algorithms that respond to changes in streams of data. An example of adaptation in our social world is the response to the COVID-19 pandemic. As a result of COVID-19, there was a reconfiguration of workplaces (from an office to home), a shift in the nature of interactions (online as opposed to in-person), a change in behaviour (distancing, wearing masks etc.), and a rerouting of supply chains.

The RDD System is adaptive. NREGA, by design, was created to be a demand-responsive program, that is, workdays and NREGA enrolment should expand in response to rural distress and a dearth in employment, and contract in response to a surge in income-generating activities in rural areas. During the COVID-19 pandemic, for example, NREGA expanded at an unprecedented scale to account for the surge in demand for employment. In June 2020, just three months after the COVID-19 pandemic broke out, NREGA was providing more employment than it did over the previous 14 years of its existence (Narayanan, 2020). This demand-responsiveness is characteristic of an adaptive system. The SHGs too reoriented their activities in response to the coronavirus pandemic. Rural women once engaged in stitching school uniforms and making soaps for government schools repurposed their activities to sew masks and produce handwash and sanitisers for the pandemic. Overall, 20 thousand SHGs in India, across 27 states produced 19 million masks, more than 100,000 litres of sanitisers and nearly 50,000 litres of hand wash (*In India, Women's Self-Help Groups Combat the COVID-19 (Coronavirus) Pandemic*, 2020). In response to disruption in the food supply, SHGs set up community kitchens, to feed abandoned migrant workers.

Figure 2: Features of a Complex Adaptive System

Openess	A feature of a system which cannot be truly isolated from its boundary because it is influenced not only by internal forces within the system but also by external forces of the overarching environment
Nestedness	A feature of a complex system in which the elements or subsystems within the system are themselves complex systems
Non-Linearity	A feature of a system in which a slight change in one part of a system can have a large impact on the overall system, or vice-versa
Dynamism	A feature of a system which changes and evolves over time, such that, it has a trajectory, a history, and is sensitive to initial conditions
Unpredictability	A feature of a system in which long-term predictions are almost impossible
Cyclicality (Feedback Loops)	A feature of a system in which effects on elements of a system cause the element itself to be altered
Self-Organization	A feature of a system where no entity or authority has absolute control, in which striking manifestations of overall order arise from local interactions between parts of an initially disorderly system
Emergence	A feature of a system that develops its own behaviours and patterns, which cannot be deduced from its parts
Adaptation	A feature of a system to respond to changes and learn from experiences, interactions, patterns, and dependencies within and outside the system

II. Tensions between top-down (macro) and bottom-up (micro) forces

So far, we have discussed, features of complex systems—openness, nestedness, non-linearity, dynamism, unpredictability, cyclicality, self-organization, and emergence — using the RDD System throughout as a guiding example. We also introduced complex adaptive systems, a class of complex systems with the capacity for adaptation.

How do complex adaptive systems work and function? Let us use the RDD system to illustrate, and start at the micro-level, with a key grassroots actor, a poor rural woman.

Poor rural women, like any other actor in the RDD system, can adapt, that is, they can adjust their own state or way of functioning given some change in the environment (Complex Adaptive Systems, 2015). Adaptation could also come from many other actors in the system, including state workers, civil society etc. In the example in the previous section, we discussed how in response to COVID-19, poor rural women in SHGs responded by helping with the production of personal protective equipment (PPE), such as masks, and the building of community kitchens. Poor rural women, in this case, responded to local information; an influx of migrant workers to villages in need of a job and food, and the shortage of masks when the spread of the coronavirus was making people severely ill. However, the poor rural women did not have a global vision of the nature of the virus, policy actions (for example, lockdowns) and the consequences (for example, disruption in transportation and thus food supply) causing the exodus they witnessed.

When migrant workers returned without work and income to support their families, each rural woman and their respective households together had to choose how to act, for example, whether or not to join current efforts or start new initiatives to help in PPE production and support the work in community kitchens. In the language of complexity, actors have the choice or autonomy to synchronize or de-synchronize their states with that of other actors (Complex Adaptive Systems, 2015). Put simply there is a choice about whether to cooperate or compete, a choice which is resolved based on an individual or group's assessment of the desirability of each choice (Complex Adaptive Systems, 2015). In this case, with public health and supply chains floundering, the incentives to cooperate were high (unlike in a zero-sum game scenario where the incentive to compete is higher). An initiative by one poor rural woman who convinced her SHG group then serves as an attractor that encourages women in other SHGs to start similar initiatives. Soon the VO (made of 10-15 SHGs) awaken to these efforts. This puts a positive feedback loop in motion, where what one woman does influences another, causing more women to participate and more women to coordinate their efforts in their VOs. As more women join efforts, the incentive to join is stronger and people realize the gains of coordination, and the positive loop is amplified. In this fashion, a small event, like a poor rural woman deciding to learn how to make masks, results in a chain reaction that cascades into her VO, her CLF, and then soon CLFs in other parts of the block. Similar chain reactions are sparked in other parts of the country, and soon this results in SHG women around the country participating in what looks like a cohesive and coordinated community-led COVID-19 response effort.

This process of synchronization is a demonstration of self-organization, in which interactions of individual actors (in this case poor rural women) resulted in a striking macro phenomenon— the production of 19 million masks by SHG women, more than 100,000 litres of sanitisers and nearly 50,000 litres of hand wash across 20 thousand SHGs in 27 states in India just a few months after the first case in India (*In India, Women's Self-Help Groups Combat the COVID-19 (Coronavirus) Pandemic*, 2020). This macro phenomenon (Complex Adaptive Systems, 2015) could not have been inferred from the behaviours of the individual poor rural women in isolation. Notice while there is remarkable coordination of COVID-19 efforts among SHG groups in India, the control and regulation of these efforts is highly distributed, as opposed to concentrated with one authority. No one women or VO leader or CLF leader has complete information about the COVID-19 response efforts or has the power to change the course of the efforts on strength of their own will. It is this decentralized and distributed nature of the RDD System (and complex systems in general) that makes them able to adapt and remain robust to large disturbances such as the COVID-19 pandemic.

In this case, many micro-level actions amounted to a macro pattern. This organizational pattern has a core tension that often characterizes complex systems (Complex Adaptive Systems, 2015) — between the macro-level organization (the pan-India coordinated community-led COVID response effort) and the micro-level actors (the poor rural women). The nature and success of the new organization crucially hinge upon how these core tensions are resolved. Is the community-led COVID response effort in India led by the interests of poor rural women, at the expense of a well-coordinated organization? Or conversely, are the interests of the community-led COVID response machinery being favoured at the expense of the poor rural women? Is there a resolution to these tensions between the top-down and bottom-up forces?

While the RDD System and its COVID response is only one illustrative example, we often see these tensions in complex systems generally, and in development and economics more specifically. Social, political, historical, and economic forces can coalesce to create a free-market economy, driven primarily by the interests of individual actors in a bottom-up fashion at the expense of societal goals. On the other hand, forces can combine to create a more centrally planned top-down economy that privileges society (or sometimes the elite high society) at that the expense of individual freedoms. Of course, in many instances, we have a combination (with India as a pertinent example of top-down centralized bureaucracy on one hand, and, at least in conception, a bottom-up Panchayati Raj system as envisioned by Gandhi). Nonetheless, this tension is at the core of political, economic, and ideological fights between the left and right.

Complex adaptive systems, because of their inherent characteristics, tend to have tensions between the top-down and the bottom-up systems (Complex Adaptive Systems, 2015). This is an important implication that we will revisit as we understand how complex adaptive systems and development are related.

2.2. Complex Adaptive Systems and Development

2.2.1. Development as a Complex Adaptive System

> "By development I mean the movement upward of the entire **social system**, and I believe this is the only logically tenable definition. The social system encloses, besides the so-called economic factors ... the distribution of power in society; and more generally economic, social, and political stratification; broadly speaking, institutions and attitudes ... The dynamics of the system are determined by the fact that among all the endogenous conditions there is circular causation, implying that if one changes, others will change in response, and those secondary changes in their turn cause new changes all around, and so forth."

(Myrdal,1974, p.729-730)

What do we mean when we say development? For many economists, development is the shorthand for 'economic development' and is even sometimes synonymous with 'economic growth' (broadly defined as the rate of increase in income per capita). In this worldview, the main measure of development is income, and poverty is the lack of sufficient income to consume the bundle of goods that society deems the minimum acceptable. Amartya Sen, and his capabilities approach, expanded the notion of development (Sen, 1999). In this world view, development means the expansion of the possibilities all individuals have to pursue a life of their choosing. This is influenced by many factors: health, education, engagement in the community, dignity (or the ability to go about without shame) as well as the material standard of living. It follows directly from this conception of development that poverty, too, encompasses a range of deprivations. This approach inspired the UN Human Development Index and multidimensional indices of poverty. Sen argued that development was inextricably linked to five freedoms (a synonym for capabilities): political freedoms, economic facilities, social opportunities, transparency guarantees, and protective security. In this world view, development is freedom, and poverty is suppression (Sen, 2014).

In all these theorizations of development in the field of economics, it is easy to overlook the origins of the word. The etymology of the word refers to a "*gradual unfolding*" or "*an internal process of expanding and growing*" or "*advancement through the progressive stage*." This points to a process of transformation and change. A definition of development that acknowledges its etymology, and the progression of its meaning in the field of economics is the following: *development is the gradual process of change that leads to improvements in income, capabilities, or freedoms (political freedoms, economic facilities, social opportunities, transparency guarantees, and protective security)*.

This definition implies that development is not the mere improvement of income and capabilities, but the *process or mechanism of change* that leads to this improvement. The process or mechanism of change is the 'social system', which is eloquently described in the quote above from the Swedish economist, Gunnar Myrdal (a Nobel-prize-winning pioneer of development economics). The social system is a complex adaptive system, comprising of social, cultural, political, and economic subsystems within it. In other words, *development refers to the capacity of our social system, a complex adaptive system, to improve the capabilities, or freedoms, of individuals, including within this, the benefits of increased incomes* (Barder, 2012a; Barder, 2012b; Beinhocker 2006).

The fact that the social system is a complex adaptive system has important implications for development. From the definition above, it naturally follows that nearly any problem or challenge in development has complex elements and features of a complex adaptive system.[8] This is not difficult to illustrate. Take any problem in development. That problem, however simplistic, when unpacked, is likely to involve multiple actors, with different reciprocal obligations to each other. Moreover, the solution to that problem will likely involve many contingencies and conditions. This always holds because no problem in the 'social system' exists in a vacuum.

Say we need to build a dam, to store water for the harsh summer. This may seem on the surface, less complex, because the mechanics of how to build the dam are known, and experts can build a dam in a location. However, while the building of the dam itself is relatively simple, everything that surrounds the building of the dam is complex. The dam is built in a location, and that location has people, a history, and an environment, all of which might be affected. There are multiple stakeholders involved (policymakers, environment activists, engineers, local inhabitants, civil society actors etc) each with their own incentives.

---

[8] The idea of complexity in development is gaining prominence although it is not yet incorporated fully in mainstream economics and social sciences. Bamberger et al., 2015 explore five dimensions of complexity in development—the nature of the system (this includes social, cultural, historical etc. forces), the intervention, the institutions and stakeholders, causality and change, and evaluation. According to them, complexity in development increases with scale and geographic spread, the diversity of the intended beneficiaries, the number and range of program components, and the number of stakeholders involved (organizational complexity).

Figure 3: The social system as a complex adaptive system

Is the social system a complex adaptive system?

Most definitely. It consists of multiple actors, entities, and processes, such as people, firms, institutions, and regulations, that are constantly interacting with each other. More concretely the social system is:

1. ***Open*** because it is influenced by forces outside of the system, for example, the ecological environment, climate, and the planet.
2. ***Nested*** because it has several complex adaptive subsystems, among them the economic subsystem* or the economy (which includes the government, regulatory bodies, ministry of finance, reserve banks, commercial banks etc.).
3. ***Non-linear*** because small perturbations can lead to large events that majorly disrupt the social system. A slight increase in global temperature (resulting in extreme weather) or the introduction of a virus (resulting in an influenza or coronavirus pandemic) or an assassination (resulting in World War I) all has large impacts on social system.
4. ***Dynamic*** because it has a history and trajectory. Over time, the social system has witnessed widespread changes in the exchange of information, the modes of communication, the systems of political organization, the rise and fall of nation-states, and the balance of power.
5. ***Unpredictable*** as evidenced, for example, by the poor record of economists in predicting recessions (An et al., 2018; Cedric & McIntyre , 2019 ;"GDP Predictions Are Reliable Only in the Short Term," 2018) and our inability to anticipate social movements. The social system contains various feedback loops, for example, the positive feedback loop that resulted on the internet and social media revolution or the rise of currency as an accepted means of exchange.
6. ***Self-organizing*** because no one actor, entity, or process has complete control of the state of the global order, and yet we see striking examples of organization and global effort (e.g., the formation of the United Nations, and World Bank).
7. ***Emergent*** because no combination of actors, entities, or processes of in a social system, can depict an accurate picture of the whole social system.
8. ***Adaptive*** because the social system can respond to events and crises, and learn from the past.

*The idea of the economy as a complex adaptive system gained prominence in the mid-1990s and early to late 2000s, largely led by the Santa Fe Institute. The movement began questioning the idea of the economy as a static equilibrium, instead proposing more complex and dynamic models that incorporated positive feedback loops of innovation. The term 'Complexity Economics' emerged from this movement (Arthur, 2009 ; Beinhocker 2006)

Let's take another example. Consider a leak in the thatched roof of a remote tribal household, say, due to the growth of algae and fungus during the monsoon season. The leak has caused water-borne and fungal diseases to spread in the household. Some elements of this constitute a simple problem, with many known solutions, among them, the use of algicide. However, let us unpack this problem further. Given the remote nature of the village, the nearest market that has the algaecide is very far. Moreover, heavy torrential rains have also damaged the roads. In the market, algaecide is short in supply due to a strike among workers in the factory producing it. Worse, the monsoon has wiped out the farmer's crops, so they are in debt. Finally, some members of the tribal community might be opposed to the use of chemicals for this purpose. This is a fictitious example, but one can see how this problem can become increasingly complex. Even if the farmer were to get algaecide, it would mean that many processes and systems were functioning well, including the village economy, the roads, and the production in factories of algaecide, to name a few.

Note while every development problem has complex elements, and features of complex adaptive systems, not every single part of the problem is complex. The problem would have some elements that are simple and others that are complex. The complex elements typically come from the systemic nature of the problem that naturally arises in development. For example, in the thatched roof example, the problem of how to fix the thatched roof is simple, but how to do it for a tribal family in a remote village is complex. The part that is complex comes from the complex adaptive system the tribal family is a part of. Similarly, the problem of how to build a dam is complicated (the engineering elements are known and predictable), but how to get the dam built, and the social, political, and economic tensions are complex. Again, the complexity comes from the multiple stakeholders and their interactions, which are intrinsic to the development process.

2.2.2. Innovation and Scaling as Complex Adaptive Systems

One of the many attributes of complex adaptive systems that development exhibits is a tension between top-down (macro) and bottom-up (micro) forces (see Section 2.1.3, Part II). Isabel Guerrero calls this the *missing middle* problem in development — top-down initiatives and policy advice by large multilaterals, governments, and aid agencies, rarely trickle down to the poor, while bottom-up initiatives and innovations by grassroots organizations that help the poor, rarely scale (Guerrero & Walton, 2015). For the poor to become architects of their own destiny, successful innovations and practices from the base of the pyramid need to achieve significant scale. This brings us to a central puzzle in development—how can we foster innovations at the base of the pyramid and how can we ensure these local innovations scale? A key to a better understanding of the puzzle lies in the recognition that innovation and scaling, like development, also exhibit features of a complex adaptive system.

Let us clarify some concepts first. Creativity refers to the creation of original ideas by individuals or by groups working together (Amabile, 1988). Innovation—a closely related concept—builds onto the notion of creativity—it is the effective implementation and adaptation of creative ideas (Farr, 1990). Scaling—an extension of innovation—involves expanding the reach of an innovation to a substantial part of the relevant population —allowing it to affect more and more lives and livelihoods. This may involve extending an innovation in one part of an organization to other parts, the growth of the organization itself, or replication of the innovation in other organizations(Cooley & Guerrero, 2016).

Creativity, innovation, and scaling are often conceptualized as complex adaptive systems (Turner & Baker, 2020). Seemingly small bursts of creativity emanating in an individual or small group can catalyse self-reinforcing positive feedback loops. Other groups and organizations slowly recognize the value of an idea, launching myriad experiments to find ways to implement and adapt the idea. People iterate on many designs and prototypes of the idea to develop a proof-of-concept, that in turn can be implemented typically in one setting—this we call an innovation. The innovation is a product of several refinements in design that culminates in a more improved and feasible design and then serves as the basis of subsequent innovation, including on designs and processes to adapt to different settings. What we refer to as scaling is effectively the aggregation of feedback loops on innovations that make them reach a large portion of the intended beneficiaries. In this ideal sequence, there is an emphasis on learning and adaptation throughout. There is learning and adaptation in (1) turning an idea into a proof-of-concept, (2) transforming a proof-of-concept into an innovation, and (3) expanding the innovation to reach scale.

Most innovations across disciplines exhibit the self-organizing and emergent properties of complex adaptive systems. This is vividly the case for Open-Source Software in which the source code is made freely available for modification and distribution. This decentralized model of software development encourages open collaboration that serves as a container of positive feedback loops that spark innovation (Yilmaz, 2008). But it also applies to innovations that occur within organizations, or by actors in the grassroots.

The cycle of creativity, innovation, and scaling applies to the example of women's self-help groups, described before. The idea was simple—women's self-help groups can help alleviate poverty through multiple paths: enabling cheaper access to credit, facilitating livelihood creation in rural areas, and fostering women's empowerment, through increased organizational capacity and expanded critical consciousness. As noted above, the core approach started outside government, originating in MYRADA and SEWA. Soon a self-help group movement came into existence with successful implementations of SHGs at the state level that went to scale initially in Andhra Pradesh, Kerala, and Tamil Nadu, each with unique characteristics (Deshpande, 2022). These successful implementations (i.e., innovations) engendered more such programs, for example, Jeevika in Bihar, inspired by the Andhra model and launched at a national scale with the NRLM. The state-level implementation of NRLM is now often initiated with representatives of Andhra Pradesh, who train women in villages to set up their own self-help groups (this is a common story in Madhya Pradesh's Jhabua district). This is an ongoing process; there are many aspects of SHG functioning that need to be refined and tweaked, a process of experimentation and prototyping that is ongoing in several states in India today.

Innovation and scaling are central to development, as they are part of the processes that turn ideas to alleviate poverty into solutions that transform lives and livelihoods for millions. Anyone interested in development would be concerned with two key questions— (1) How can we enable and foster innovations that tackle development challenges?, and (2) What would make these innovations scale, so their impact is widespread and felt by many around the world? This second question of scaling is arguably even more challenging. There are often many extraordinary local initiatives, that change lives for the people in their vicinity but rarely scale.

2.3. Applications of Complexity Theory to Evaluations in Development

In the previous section, we argued that nearly all challenges in development, especially, challenges around scaling and innovation are inherently *complex*, or more formally in the context of complexity theory, have features of a complex adaptive system (see Sections 2.1.2 and Section 2.1.3, Part I for all the features of a complex adaptive system). This assertion implies that addressing challenges in development around innovation and scaling involves system-level processes of change. Standard impact evaluations used commonly in development (for example randomized control trials) are good for statistical assessment of the impact of an intervention in a particular context but are not well designed for these types of system-level change processes, and typically do not take a complexity theory lens on development (Bamberger et al., 2015; Byrne, 2013; Forss et al., 2011; Stern, 2012).

What does complexity theory imply for evaluations in development? What lessons from our understanding of complexity in development are central to how evaluations should be designed and run? What does an evaluation that applies the concepts of complexity theory, and is truly designed for supporting system-level transformation look like? How does a complexity-based evaluation relate to existing approaches to evaluation, including RCTs and similar techniques? What methods, techniques and evaluations currently exist to address complexity? This section aims to delve into all these questions, and more.

2.3.1. Implications and Challenges of incorporating Complexity Theory in Evaluations

The fact that nearly every problem in development resides in a complex adaptive system has important implications that influence how we evaluate programs and interventions. Below are some of the main challenges and implications of incorporating the ideas of complexity theory in evaluations in development. In describing the challenges, each part also outlines what an evaluation that addresses these challenges would look like and what methods and techniques can be used to address these challenges.

I. An evaluation cannot simply evaluate the parts of a system to understand the whole

Many standard impact evaluations in development, such as RCTs, have an implicit assumption that by understanding the inner workings and effectiveness of a *part* of a complex problem (typically a specific intervention) we can gain insight into the problem as a *whole*, and effect the transformation we want to see in terms of outcomes. What complexity theory has shown us is that to understand any problem in development, it would be reductionist to only look at parts of a system because the whole is *different* from the sum of its parts (Turner & Baker, 2019).

An evaluation that truly recognizes the centrality of complexity theory in development would aim to (1) understand a specific part of a system (say an intervention), as in a standard impact evaluation and (2) understand how that part (the intervention of interest) interacts with other parts of the system. This would provide a more comprehensive and complete understanding of the intervention's influences, and feedback loops, as part of a broader assessment of the ability of the intervention to induce change in the system.

Information on the causal influence on an outcome of interest is of course important information and RCTs and similar techniques are powerful tools for this, though not the only methods. Rather, from the perspective of complexity theory, understanding whether an intervention had a specific causal impact on an outcome, is one step in a two-step process, in which understanding how the intervention interacts with other parts of a system is equally important. Both steps are complementary and incomplete without one another.

The peril of _only_ assessing whether an intervention has a causal impact in the specific context[9] is that it may give us a distorted picture of the ability of an intervention to induce transformative change within the system. An intervention may work in isolation but might be misaligned with the incentives of actors responsible for implementing it, crowd out the positive impact of other interventions aimed at addressing the same problem, or not be feasible given the resources, abilities, and capacities of the system. For all these, it is important to understand how an intervention interacts with other parts of the system.

On the other hand, the peril of _only_ assessing how an intervention interacts with the system is that we would have a poor understanding of whether the intervention itself works in the first place. An intervention may be incentive compatible for all actors in the system, not crowd out other interventions addressing the same problem, may be politically favourable, and be feasible within the system, but may be ineffective and flawed in its intrinsic design. We need both steps to get a complete understanding. The steps are individually necessary but only jointly sufficient to create meaningful change.

While this is ostensibly "just" an extension to a standard impact evaluation, it is in some way still a radical insight. It goes against the conventional wisdom deeply embedded in our thinking that to solve a complex problem, one can break it down into smaller more accessible problems (these smaller more accessible problems are typically the ones evaluated by an RCT). Even the etymology of the word *analysis* refers to "*the resolution of anything complex into simple elements.*" Economists often see themselves as analysts, in that very sense, simplifying complexity through models. Dani Rodrik and many other economists often give the analogy of economic models as maps (Rodrik, 2015), pointing out how the abstraction from reality is precisely what makes maps so useful and informative (a truly accurate map is a replica of the world, and thus practically useless). The challenge, Rodrik says, is making sure the *right* map is used for the *right* problem, as opposed to focusing instead largely on building maps. From a complexity perspective, this is incomplete. While a model, or map, can help thinking, and approximate observed patterns of behaviour, it will be insufficient for interpreting a development problem as it unfolds, precisely because the economy is a complex adaptive system that exhibits openness, nestedness, non-linearity, dynamism, unpredictability, cyclicality, self-organization, and emergence. In other words, there is no *right* map for a problem.

Instead of only looking at problems from the narrow prism of smaller parts or abstractions of parts of the system, (that seek to explain the whole)[10], complexity theory urges us to also examine problems from the purview of systems change, specifically, which parts of the system can be activated to ignite a series of positive

---

[9] Unfortunately, economics and the social sciences continue to privilege the analysis of a particular intervention much more than a systematic examination of how the intervention interacts in the system

[10] Many methodologies, aside from RCTs, tend to fall into the trap of using parts to explain the whole. Growth Diagnostics (Hausmann et al. 2008), aims to find the binding constraint to economic growth, the idea being that the removal of a small number of binding constraints will unlock spurts of economic growth. This emerges from the view that economic growth is the result of a constrained optimization problem. From a complexity standpoint, there are many issues— (1) the economy, being a complex adaptive system cannot be described by a constrained optimization problem, (2) one path of a decision tree (a part) rarely alone explains the root of the problem (the whole), that is, the paths of the decision tree may not be independent, so forces that explain one cause could also explain another; (3) instead of a small number binding constraints causing development failure, it could be multiple causes acting simultaneously, (4) a constraint of today may not be the constraints of tomorrow, and (5) relaxing one constraint may hamper progress in another, because they are interrelated (Felipe et al., 2011).

feedback loops that eventually put us on a trajectory of better growth, capabilities, and freedoms. From this perspective, the task of modelling becomes one of updating and capturing change, rather than seeking to reduce complexity to separable, constituent parts, which is often the goal of traditional economic analysis.

The idea that complexity can be broken down into simple parts is widespread. What complexity theory shows us is exactly the opposite, a system is complex because it cannot be broken down into simpler parts.[11] This calls for analysing each part, coupled with an examination of how different parts interact with one another.[12]

The techniques used to understand the first step. (i.e., understanding a specific part of a system), are plenty and are widely used in standard impact evaluations. These include RCTs, quasi-experimental methods etc. There are many techniques, both qualitative and quantitative, that help us understand the second step (how an intervention interacts with other parts of a system). These are discussed further in Section 3.2.2. Qualitative methods include, among others, (1) case studies of an intervention, involving interviews with different actors and stakeholders of an intervention to develop system descriptions (Paparini et al, 2021; Widner et al., 2022), (2) participatory actor mapping exercises (Gopal & Clarke, n.d.), in which actors come together and map reciprocal obligations to each other, and (3) system diagnostics (Waddell, 2016), to identify enablers and resistors of change. Quantitative methods include, among others (1) system dynamics (Forrester, 1999), developed at MIT, which undertakes mathematical modelling of several feedback loops, (2) agent-based modelling (Gallegati & Kirman, 2012), which simulates actions and interactions of autonomous agents, and (3) social network analysis (Freeman, 2006), which uses graph theory to model social connections (Walton, 2014).

II. An evaluation cannot ignore history and context

Almost everything in development has an evolution, and the past and present have a bearing on how the future unfolds, and lead to path dependence of a system. Standard impact evaluations typically do not delve into the trajectory of an intervention and the trajectory of those actors and entities that are affected by or influencing the intervention. Yet, the life stories of the participants in the treatment and control groups are typically profoundly affected by history, the current context, and relationships between individuals with other groups and the state. This is also important when considering how to scale an intervention because the context, history and the environment shift as one moves into new territories, and the intervention's implementation needs to change to adapt to these altered realities.

Case studies and ethnographic studies (Anderson et al., 2005) typically invest time in developing a detailed understanding of system history and initial conditions at the time of an invention (Walton, 2014).

[11] Zimmerman et al 2011 called this a conflict between complexity reduction and complexity absorption. The paper concludes that" if *the nature of the challenge is such that complexity (structural, cognitive, or social) cannot be reduced without reducing the resonance and utility for the users of an evaluation, then absorbing complexity will be more effective in the long run*." The paper highlights how attempts to reduce complexity in healthcare led to negative long-term effects.

[12] Bamberger et al., 2015 propose a useful method to do this—an Unpacking-Reassembling approach, in which one maps the levels of the system, unpacks the system to understand each constituent part, then reassembles the system, this time, with more careful attention to links between parts, and finally goes back to assess the big picture

III. Cause and effect relationships are rarely linear or uni-directional

Causality rarely runs only in one direction. David Byrne in his paper, *Evaluating Complex Social Interventions in a Complex World* (Byrne, 2013), put it rather eloquently,

> "So, parts (of a system, such as an intervention) have causal implications for the whole, interactions among parts have causal implications for the whole, parts have causal implications for each other, and the whole has causal implications for parts."

Standard impact evaluations by design tend to focus on trying to disentangle the singular impact of a particular intervention (e.g., women's collectives) on the outcome of interest (e.g., agency, say for sake of illustration, measured by women's household decision-making), keeping all other factors constant.

Complexity theory suggests that one intervention is rarely the only or dominant cause of a change in an outcome (Byrne, 2013). Often what matters is not the intervention itself, but instead, how the intervention works in relation to the systems and sub-systems that surround the outcome of interest. To illustrate, women's collectives are unlikely to be the primary cause of increased women's household decision-making. The success of women's collectives in moving the needle in household decision-making is largely contingent upon how the women's collective system functions, and how that system interacts with the village governance system, the patriarchy (itself a system of society), the larger social fabric of the family, and the politics of the leadership at the highest levels.

This is why asking different stakeholders about an intervention can lead to vastly different judgements of 'what worked', as key actors have different assessments of influences on outcomes, and indeed different values. For example, a male head of the village council, might not think of women's household decision-making as an important value, and thus might think women's collectives worked because they bring more credit to the household as opposed to because it means women make more decisions.

Finally, rarely is the intervention of interest the only intervention operating and interacting with the outcome, which makes it challenging to identify the effects of one intervention from another (Walton, 2014). Within a system, many other interventions are also attempting to improve women's agency (including initiatives to get women to work, cash transfers to women, and counselling to women on empowerment)

Non-linear multi-directional relationships are indeed a challenge for an evaluator to comprehend. But to reduce these to linear, uni-directional causal relationships would be to misunderstand how development works and takes place. Many techniques are built to address non-linear, multi-directional relationships. These include among others, case studies, process tracing, outcome harvesting (Wilson-Grau & Britt, 2012), and Qualitative Comparison Analysis (Blackman et al., 2013).

IV. The timing and nature of impacts from an intervention are not obvious or known
Another challenge of the non-linear interactions in a complex adaptive system is that it is impossible to accurately predict the timing and nature of the impacts arising from the intervention (Walton, 2014). It is unclear whether women's collectives' impact on agency would materialize after a few years, in decades, or after generations. In addition, it is unclear how the agency would manifest itself. Would an increase in agency be reflected in greater labour force participation, increased mobility or greater influence in household decision-making?

An evaluation that understands that the timing and nature of impacts are not known would have more varied and frequent measurements over a longer time frame, working closely alongside implementation teams in the development of a program or intervention, recognizing that impacts may show up at a later time.

2.3.2. Current evaluation Approaches Designed for Complexity

Standard impact evaluations are ill-designed for addressing complexity for the reasons discussed; they evaluate parts of a system to infer about the whole, they typically underemphasize history and context (and in general sensitivity to initial conditions), they often assume linear and uni-directional cause and effect relationships, and expect the timing and nature of impacts to be detectable within the evaluative frame.

By contrast, there are many techniques, some old and well-known, such as case studies, and some relatively new in their use in economic and development contexts, such as process tracing, outcome harvesting, and qualitative comparison analysis, which are more suited to the analysis of complex systems.

However, evaluations based on the tenets of complexity theory in development are few. One example is Pawson and Tilley's Realist Evaluation, whose philosophical underpinnings are relatively close to complexity theory, in the particular sense, that it rejects the idea that identifying linear causal mechanisms are central to an analysis of program and policy effectiveness. More important to a realist evaluation is considering the theoretical mechanisms hypothesized and the socio-historical context an intervention resides in (Pawson & Tilley, 1997). In other words, a realist evaluation attempts at understanding in a particular context, what generative mechanisms or motivations for the actors result in particular outcomes, sometimes referred to as the context-mechanism-outcome (CMO) framework (*Realist Evaluation*, 2014). While there is an acknowledgement that initial conditions matter and that interventions affect multiple levels in the system, practically, the evaluation approach is not very precise about the tools and techniques to use to apply the CMO framework. Moreover, while a realist evaluation is method neutral, there is little guidance on which among the gamut of methods is best suited to attempt to answer what works for an intervention, for whom and in what context. In addition, relative to other evaluation methods, there are few examples of good realist evaluations.

Perhaps the most direct application of complexity theory in evaluations in development is Michael Quinn Paton's Developmental Evaluation. Developmental Evaluation, in the words of Paton, "*supports innovation development to guide adaptation to emergent and dynamic realities in complex environments.*" (Patton, 2010) The approach of a developmental evaluation is similar to the role of research and development in the private sector (*Developmental Evaluation*, 2011), in that it emphasizes gathering near real-time data as feedback for continuous improvements, adaptation and experimentation. Some tools used in a development evaluation include evaluative questioning (asking What?, then So What? And then Now What?), network mapping, logic models and theories of change, simulations, and appreciative inquiry (Gamble, 2008). Finally, in a developmental evaluation, an evaluator is embedded within an organization where innovation is taking place, as opposed to an outsider who gives an independent assessment. The idea is to support social innovations within an organization and co-create, in an iterative fashion, an improved solution that is consistent with the values of the innovator. While developmental evaluations are undeniably designed for complex environments, they are restricted to innovation within an organization or set of organizations. The adoption or replication of an innovation in other contexts, and how the innovation can scale, is outside the purview of a developmental evaluation (Gamble, 2008).

Both a realist evaluation and a developmental evaluation have their relative strengths and weaknesses. Realist Evaluation emphasizes the importance of past and initial conditions (i.e., context and history), and generative mechanisms (i.e. theory) in the realization of an outcome. It does not necessarily give a clear roadmap of the methods and techniques to measure these initial conditions and methods. Developmental Evaluation focuses on innovation in complex environments. It is not designed to necessarily address questions of the socio-historical context that are addressed in a realist evaluation or questions of scale (via the government or the market).

Yet any account of transformation and systemic change should be aware of initial and past conditions, the role of innovation, and the role of scaling those innovations. An Adaptive Evaluation, which we formally introduce in the following section, aims to do just that. It has two main contributions to the existing evaluations in development that are based on complexity. First, as implied above, it combines (1) an evaluation framework for an in-depth understanding of the role of context, history, and theory in outcomes (similar to a realist evaluation), with (2) an evaluation framework to support innovation (involving rapid measurement, prototyping and experimentation (as in a developmental evaluation), and (3) an evaluation framework to support scaling. An Adaptive Evaluation is in this way more holistic, in that it spans the spectrum of key issues in development from a complexity lens, but this means it is also much more ambitious in its scope. Second, an Adaptive Evaluation offers a clear strategy, not only of techniques and methods that can be used in each of the three evaluation frameworks but also of under what conditions is it best to use these techniques and in what sequence. This is important, as one of the drawbacks of an evaluation based on complexity is how to implement such an evaluation from a practical standpoint.

Overall, the Adaptive Evaluation approach has similar philosophical underpinnings to a realist evaluation, in terms of understanding the role of context, history and theory but the framework to assess these is different. Likewise, the Adaptive Evaluation framework for innovation has similar philosophical underpinnings to a developmental evaluation but differs in the application and techniques used. The evaluation framework for scaling is not inspired by either the realist or the developmental evaluations.

Finally, as mentioned in the introduction, we use the word '*adaptive*' in Adaptive Evaluation for four main reasons. First, the *adaptive* points to a recognition that development, scaling, and innovation all share features of complex adaptive systems (as outlined in Section 2.2). Second, adaptive refers to the emphasis on continuous learning and improvement to *adapt* intervention design to the evolving needs of the end beneficiaries and to *adapt* to new contexts and settings as the intervention scales. Third, the evaluation itself is adaptive, in that the methods and techniques used are intended to adapt to the nature of the task (innovation or scaling), and the complexity of the system and environment. Finally, we also use adaptive because of its links to adaptive leadership and adaptive implementation methods. Note that this is a different usage of "adaptive" to its recent usage in the econometrics literature, which refers to the (interesting and useful) approach to treatment assignment and/or confidence intervals for adaptive experimental designs[13](Kasy & Sautmann, 2021; Hadad et al., 2021). The use of adaptive in an Adaptive Evaluation is broader, and not confined to RCTs.

3. Introducing Adaptive Evaluation

"... a practice of social and historical explanation, sensitive to structure but aware of contingency, is not yet at hand. We must build it as we go along, by reconstructing the available tools of social science and social theory. Its absence denies us a credible account of how transformation happens."

(Unger,1998, p.26)

In the quote above, Unger highlights the need for social science to provide a 'credible' account of how transformation happens. He emphasizes that to be able to truly inform our actions towards the future we seek, we need to be able to understand how things became as they are. We need to understand how transformation in development happens to inform our policy actions about how to alleviate poverty. And we need to understand how extraordinary local initiatives to alleviate poverty were able to scale to understand how to co-create a world in which all local grassroots innovations achieve global scale.

Byrne argues complexity theory offers the foundation of the kind of social science Unger eloquently describes because it is inherently dynamic—it concerns itself with the trajectory of change in a system (Byrne, 2005). An Adaptive Evaluation is an attempt to use complexity theory to (1) build a credible explanation of how transformations take place, and (2) use it to inform actions to foster innovations in development and take them to scale. The hope is that an Adaptive evaluation approach can build the tools and theory to understand how transformations happen, serving as a foundation for social action towards a future in which deprivations are eliminated for all.

[13] Adaptive experiments are as almost as old as RCTs (Thompson, 1933), The idea of an adaptive experiment is to ensure that participants get the best possible treatment outcomes. A useful analogy is that of the 'multi-armed bandit problem' (Green, 2018). This refers to a game in which one can drop a coin in many slot machine arms, each with different pay-off rates. There is a tradeoff between, exploration— learning about different slots individually, and exploitation— the opportunity to maximize benefits by using the best arm. The same tradeoff applies to experiments, thus, an experimenter should (1) devote some share of the experiment to other treatment arms (i.e., exploration), but (2) assign most participants to the treatment arm shown to be most effective in prior waves.

This section formally introduces an Adaptive Evaluation. Section 3.1 looks to develop a framework for evaluation questions, and introduces the idea of complex *problems*, and *complex questions*— that are subject to unpredictable and constantly evolving human behaviour, systems, and contexts. Section 3.2 then discusses categories of complex questions and methodological elements of an Adaptive Evaluation.

3.1. Motivation

The aim of Section 3.1 is to put together the final pieces that allow us to define an Adaptive Evaluation. In Section 3.1.1, we first introduce David Snowden's Cynefin framework, as a useful way to characterize contexts and problems in development, that was developed from complexity theory (Snowden & Boone, 2007). This serves as a backdrop for Section 3.1.2, which (1) introduces types of evaluation questions, and (2) for each type, draws parallels with the Cynefin Framework.

3.1.1. Understanding Contexts and Problems in Development

Most evaluations attend to problems and associated questions, and thus the nature or kind of problem often informs the kind of evaluation. In this section, we use the Cynefin Framework to categorize the contexts of a problem, and then look at what this implies for evaluations.

I. *The Cynefin Framework*

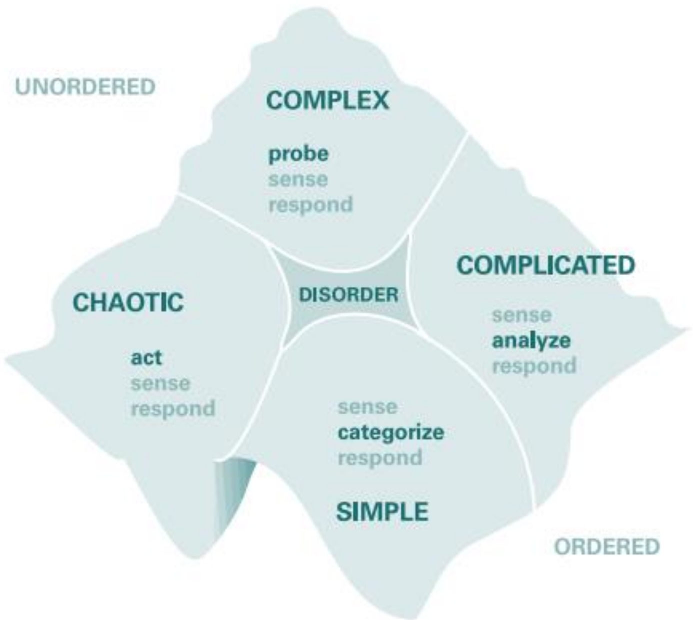
The Cynefin framework presents a novel approach to leadership and decision-making based on complexity theory (described in Sections 2.1 and 2.2). The framework[14] describes four main decision-making situations (or "domains") and how to navigate them—simple, complicated, complex, and chaotic (Snowden & Boone, 2007).[15] An illustration of the Cynefin Framework is in **Figure 4.**

In the *simple* domain, the solution of a problem is known in advance and can be repeated with the same consistent result. For 'simple' problems, what needs to be done is self-evident, and can be pulled easily from best practices or prior knowledge. In this case, the cause-and-effect relationship is clear and undisputed—the solution almost always leads to the desired outcome. To address simple problems, one needs to review the facts as they are, categorize the problem, and apply the appropriate rulebook or manual (sense-categorize-respond). The classic example of a simple problem is cooking, say for sake of illustration, a Thanksgiving Turkey, because one can follow a recipe, which works each time if followed correctly (Glouberman & Zimmerman, 2002). As the example of a recipe illustrates, while the decision-making is analytically simple, the process of getting to this "known fact" can involve an extensive journey. Snowden, in his 2007 article for the Harvard Business Review, uses the example of the loan-payment process to describe a simple problem (Snowden & Boone, 2007). An example from development is learning how to sew clothes (manually or with a sewing machine) or learning how to plough fields (with an ox or with a tractor).

[14] Cynefin comes from the Welsh word meaning 'habitat.', which signifies how we are influenced by various factors in our environment (such as our geographical setting, religion, culture, heritage etc.) that we are not always consciously aware of. It signifies how our world and worldview are shaped by inherently complex forces (Snowden & Boone, 2007).

[15] David Snowden and his colleagues in IBM's Institute of Knowledge Management developed the Cynefin framework to provide leaders making perplexing decisions a 'sense of place' from which to view their perceptions of a situation (Snowden & Boone, 2007).

Figure 4: The Cynefin Framework (Snowden & Boone, 2007)



In the complicated domain, the solution of a problem is knowable (Zimmerman et al., 2012) and at least one 'right' answer exists. Even though the solution needs to be built, it is understood to be attainable, and once built, can be repeated with the same consistent result. For 'complicated' problems, there are multiple interactions and moving parts. In practice, what needs to be done to effect changes may not be immediately obvious and will require empirical assessment of the parameters of the cause-and-effect relationship. However, these can be analysed and documented, culminating in a solution whose impact is highly predictable and reproducible. To address complicated problems, one needs to review the facts as they are, analyse a range of solutions to determine what is effective, and apply a blueprint made specifically for the problem (sense-analyse-respond). Complicated problems typically involve experts who have relevant and practical knowledge in a particular domain. A classic example of a complicated problem is sending a rocket to space because it requires a team of multidisciplinary experts to design it and make it work (Glouberman & Zimmerman, 2002). However, once achieved, it works each time. Snowden gives the example of searching for mineral deposits because it also requires a team of experts to evaluate the right spots for mining and drilling (Snowden & Boone, 2007).

In the complex domain, the solution is unknowable (Zimmerman et al., 2012) in advance. There are no 'right' answers, and a solution that worked once in one (temporal or geographic) context, may not reproduce the same result the next time. For 'complex' problems, what needs to be done is not obvious. This is particularly true of complex adaptive systems, which, as discussed above, have the features of openness, nestedness, non-linearity, dynamism, unpredictability, cyclicality, self-organization, emergence, and adaptation. In this case, the cause-and-effect relationships can only be deduced through exploratory observation of the processes in retrospect, that is, the solutions and their impacts, are inferred in hindsight through trial-and-error or other equivalent processes. To address a complex problem, one needs to experiment, fail, learn, and repeat over and over, then interpret some instructive patterns that emerge, and then respond (probe-sense-respond). Past experience, or indeed "models", can provide initial guidance, but are not the end of the story. A common example of a complex problem is raising a child, because how to raise a child is difficult to know, what works can only be deduced by trial and error, and what worked with one child may not work with another (Glouberman & Zimmerman, 2002). Development is replete with examples of complex problems – from inflation (that reduces the purchasing power of the poor) to women's empowerment, to lack of quality education for the poor, and poor health and sanitation, almost all problems in development are complex, at least until a wealth of experience and assessment has been undertaken. As mentioned before, development has features of a complex adaptive system, and so naturally has complex problems (see Section 2.2.1)

In the chaotic domain, the solution is simply unknowable—any pursuit for the 'right' answer is futile. For 'chaotic problems', what needs to be done is unknown, and there is no time to decipher information and find out what to do. In this case, cause-and-effect relationships are unclear, because they are too confounding or difficult to resolve. In the chaotic domain, what is being observed appears, at least at first, as an anomaly. To address a chaotic problem, one needs to first act (on pure instinct) to get out of the danger zone, then sense where there might be more stability, and then respond to turn the problem into a complex one. Snowden and Boone use as an example of a chaotic problem the case of a local police force facing a massacre in a remote place because what needs to be done is not knowable. There is no time to find out what needs to be done, one can only respond intuitively and wait to find the space to respond (Snowden & Boone, 2007). An example of the chaotic domain in development is the initial situation in the 2004 Tsunami in South East Asia.

The Cynefin framework was developed acknowledging that the world can be disorderly and unpredictable and that the context of a problem matters. Through the Cynefin Framework, Snowden and his colleagues emphasize that actions should be tailored to fit the complexity of the circumstances one faces. In this way, it is aligned with our discussion in Section 2.1 on complexity theory.

II. The link between problems and evaluation questions

We will use the Cynefin framework to develop a framework of evaluation questions in the next section. In this section, we will look at some important implications of the Cynefin Framework; how it relates to evaluations, and how it links to the implications of complexity theory on evaluations.

A. The nature and type of evaluation involved are different for different domains

Each domain in the Cynefin framework, except for the chaotic domain where an informed knowledge-based response is untenable, requires different kinds of evidence and has different evidence-generating mechanisms.

- In the *simple* domain, the evidence involves known facts about the world and a compilation of solutions we know work. In this domain, evidence generation involves collecting and synthesizing relevant facts, existing knowledge, and best practices.
- In the *complicated* domain, evidence is about examining and assessing whether or not a solution works, that is, does the solution cause the required improvement in the outcome of interest? In this domain, evidence on causal

relations is often generated through analysis or trials conducted on either separable parts of the system or the whole.

- In the *complex* domain, the multiple interactions among actors coupled with the dynamic and emergent features of the problem, mean that specific parts are not separable, but need to be understood within a system context. A more exploratory approach to evaluation is needed, for example with an ongoing observation of interactions and responses within the system. In this domain, evidence is often best generated using continuous experimentation involving ideation, prototyping, and testing, with feedback on designs over the process.

The chaotic domain is less interesting from an evaluation perspective and, we will omit it from the discussion in the subsequent sections.

B. It is possible to move across the domains

A feature of the Cynefin Framework is that one can transform a complex problem into a complicated one at least in a specific context, through repeated cycles of learning and experimentation, in which knowledge is gained and the parameters of the multiple relationships understood (this involves moving clockwise across the framework in **Figure 4**). Conversely, as contexts shift, mindsets change, and/or knowledge is forgotten or dissipated, a simple or complicated problem may become a complex one (moving counterclockwise across the framework in **Figure 4**).[16]

A pertinent example of transforming a complex problem into a complicated one is the story of how an innovative conditional cash transfer program in Mexico (Prospera-Progresa-Oportunidades), successfully survived over four six-year presidential terms and scaled to many countries in Latin America (Levy, 2019).[17] In 1994, Santiago Levy, then the deputy minister at the Ministry of Finance, faced a highly complex problem. Existing food subsidies (on tortillas, beans and bread) were proving to be highly ineffective because (1) they were poorly targeted—80 cents for every 1 peso spent on the food subsidy did not go to the poor, (2) the food intake was not being converted to caloric energy, because health and nutrition were deprioritized, (3) tortilla, beans and bread were not nutritionally wholesome, and (4) households that got the subsidies distributed it unequally, with later children, and particularly girls, getting a lesser share (Levy, 2019). In addition, while the supply of primary schools had increased, evidence showed children were not going to school (Levy, 2019). Finally, the 1994 macroeconomic crisis in Mexico (the Tequila crisis) meant people were likely to get poorer. There was a compelling need for investments in health and education for the poor to escape a vicious intergenerational cycle of poverty.

Levy and his team were able to make the problem less complex by experimentation and adaptation. They first piloted the idea of direct conditional cash transfers to mothers contingent on family visits to health clinics in the city of Campeche for one

[16] An example of a simple problem becoming a complex one, is that of immunization programs. For example, for diseases like Polio, it was known that the solution is to vaccinate all children. Perhaps more pertinently, it was universally agreed that children should be vaccinated to eradicate Polio, i.e., the cause-and-effect relationship between the vaccine and polio eradication was clear and undisputed. This made immunization for Polio a simple problem. In the COVID-19 pandemic, once the vaccine was available, what used to be a simple problem became a complex one. While the vaccine was known among the scientific community to be effective, it was not universally agreed that people should be vaccinated to reduce the spread of the coronavirus. How to increase take-up of the vaccine and persuade vaccine deniers, is a complex problem, when there is no consensus.

[17] Prospera was later renamed Progresa, and then Oportunidades by different political entities

year, with the Instituto Tecnológico Autónomo de México (ITAM) responsible for the impact evaluation (Levy, 2019). This was an early design of Progresa, being tried out on a small scale. The encouraging evidence from this evaluation convinced President Zedillo to support a larger rollout with 300,000 households in eight states over two years, evaluated with an RCT conducted by the International Food Policy Research Institute (IFPRI). This time school attendance and scholarships were added to the conditionality criteria for the cash transfer, The IFPRI pilot showed positive results on several dimensions, and the next phase served 1 million people. In each phase, Levy and his team were able to enhance the design of the programs, as mechanisms and effects became clearer (Levy, 2019). For example, in one pilot, his team learned that the composition of iron supplements needed to be revised to allow for better absorption of iron. In another pilot, they realized that education scholarships needed to be expanded by three years for a greater impact on high school (the evidence showed the positive impact was mostly concentrated on secondary schooling without the three-year expansion).

However, experimentation and adaption to discover designs that work were not the only sources of complexity for the problem Levy and his team faced. The broader context was of the significantly reduced fiscal space, because of the financial crisis, and a complex political situation with rising political competition after many decades of one-party rule. There was indeed a transition of the Presidency to an opposition party in 2000, the first election after Progresa had been designed and implemented. To create the fiscal space, the team used savings generated from phasing out the large-scale untargeted subsidies. This helped finance Progresa without burden on the budget (Levy, 2019). To manage the political transition, the team showcased the concrete evaluation results and skilfully managed an Inter-Development Bank loan, that provided incentives for the preservation of the initial program. (Levy, 2019). After the 2000 election, the incoming President Fox preserved the program, appropriated it under the new name of Oportunidades, and then expanded it! The relaxation of political and fiscal pressures reduced the complexity and ensured that subsequent administrations could focus on design and implementation (Levy, 2019). However, there was one dimension the program did not succeed in—it never solved the challenge of low-quality of learning outcomes. This is a further complex problem that involves many interactions between pedagogy, teacher behaviour, and the broader context of state behaviour. And as a postscript, the programme was finally cut and replaced by an alternative under the government of Andrés Manual López Obrador—on apparently ideological grounds.

C. The domains are nested

Complex problems, have within them, complicated and simple problems. This was previously shown with the example of a dam in Section 2.2.1, which had predictable, known features embedded within a complex context. The complex problem faced by Santiago Levy also shares this feature. The problem had many complex features at the outset, including what kind of program can address the intergenerational cycle of poverty and how to create political support for a conditional cash transfer, and prevent backlash from retiring the food subsidy. But this problem also has complicated elements, such as what should the composition of the nutritional supplements added to children's milk be to ensure high absorption, and how to implement a new payment system that worked with the conditionality.

D. A final note

Notice that while a complex problem can become less complex through a process of empirical exploration, this does not imply that a complex problem can initially be broken down into separable, smaller parts to be solved. Complex problems become less complex precisely through the combination of (1) understanding the functioning of parts, and (2) how the parts interact with the whole. The story of Progresa only affirms this. The complexity of the problem faced by Santiago Levy and his team was reduced by focusing on the conditional cash transfer intervention (the parts), but also on how it interacted with the budget, the politics, and the administration (the whole). It was proactively working on all these fronts that lowered the complexity. Simply improving the design through experiments without working on the politics and the budget would not have led to a sustainable solution, if this meant the programme was later abandoned.

3.1.2. A framework for Evaluation Questions in Development

In the previous sections, we were introduced to the Cynefin framework for different problems. We saw how the domains of the framework (in particular, the simple, complicated, and complex domains) have implications for evaluations.

Any evaluation is essentially a set of questions, and associated techniques, to address a problem. Questions are the 'reason for being' of an evaluation. They are the guiding light, in the sense that any technique or methodology used in an evaluation is in service of the overarching questions that define it. The characteristics of questions tackled by an evaluation are revealing about the nature of the evaluation, and what it seeks to find about the world.

In general, we suggest there are three main types of evaluation questions—simple, complicated, and complex. An evaluation may delve into all three question types or a combination of types with varying sequences over time. In this section, we describe each type of evaluation question and end with an illustrative example from the COVID-19 pandemic. **Figure 5** below summarizes the 3 broad types of evaluation questions we will discuss below.

Figure 5: A summary of the three types of evaluation questions

Simple

- Focus on facts about the ***world as it is*** and ***world as we know it***
- Involve generating **descriptive statistics** (such as a mean, variance, or correlation) or conducting a **literature review/synthesis** of past work
- Naturally arise in the **simple domain** of the Cynefin framework, but **are crucial to all domains**
- Typical questions include:
  - What is the average/percent of Z in the data?
  - What are individual views,?
  - What do we currently know about a topic of interest?
  - What are definitions, theories and concepts related to a particular issue?

Complicated

- Focus on **cause-and-effect relationships** between two or more variables
- Typically **involve lab-like settings** to isolate treatment and control and measurement over time
- Naturally arise in the **complicated domain** of the Cynefin framework, but are **crucial to the complex domain**
- Typical questions include:
  - Was X intervention effective in causing the desired Y outcomes?

Complex

- Focus on questions that depend on unpredictable and evolving
  - Human behaviour
  - Systems
  - Context
- Lab-like settings are possible in only a few instances
- Naturally arise in the **complex domain** of the Cynefin framework
- Partly **involve answering technical and descriptive questions**
- Typical questions include:
  - Does the intervention have the ability to induce the desired transformative change in the system?
  - What are enablers and resistors to transformative change
  - Does the intervention crowd out other interventions?

An evaluation may delve into all 3 question types or a combination of types with varying sequences over time

I. Simple Questions

Simple Questions focus on documenting (1) facts about *the world as it is,* and (2) the existing knowledge, theories, and concepts about the *world as we know it*. Answering simple questions about *the world as it is* typically involves generating descriptive statistics (such as the mean, variance, and correlations between variables) from a sample, to understand the sample itself and what it entails, rather than inferring about the population. Answering simple questions about *the world as we know it* often involves a literature review or synthesis of available knowledge or evidence on a given topic. Examples of simple questions are 'What is the average /percent of women who work in India?' or 'What are individual views on democracy in the US?' or 'What are consumer expectations on inflation?' or 'What are the definitions, concepts, and theories related to complex systems?'

This type of question naturally arises in the simple domain of the Cynefin framework. However, the nested nature of the Cynefin framework (see Section 3.1.1, Part II.C) means simple questions are crucial to all domains of the Cynefin framework, including complicated and complex domains. For complicated and complex problems, simple questions serve as the fundamental basis of understanding on which further analysis and investigation are undertaken. In empirical literature on development economics, the majority of the academic papers are interested in questions of causality, and yet almost all start with descriptive statistics and facts as these put the analysis of causal mechanisms in context.

Techniques and methods to answer simple questions include descriptive statistics, benchmarking, secondary or desk research, and literature reviews of established knowledge.

II. Complicated Questions

From an evaluation perspective, complicated questions focus on establishing a cause-and-effect relationship between two or more variables. In development, this typically involves testing whether an intervention 'caused' a change in the desired outcome of interest. Answering complicated questions requires lab-like settings—or "quasi-experimental" proxies for these—to isolate the treatment group (those who get the intervention) from a control group (those who don't get the intervention and serve as a comparison for the treatment group). Causality has a temporal aspect to it, in that the 'cause' must take place before the 'effect', and thus typically involves following beneficiaries over time (usually during a baseline and endline). Typical questions include "Did the women's collectives cause a measurable improvement in women's agency?" or "Did teaching at the right level in the classroom improve reading and math skills?"

Complicated questions naturally arise in the complicated domain of the Cynefin framework. The nested nature of the Cynefin framework (see Section 3.1.1, Part II.C for more details) implies that complicated questions are often relevant to the complex domain of the framework. In addition, complicated questions necessarily involve simple questions. Complicated questions often focus on how a part of the system (the intervention) is functioning, which is one important aspect of the system as a whole in the complex domain.

Techniques and methods to answer complicated questions are well-known and include RCTs or quasi-experiments or constrained optimization, among others.

III. Complex Questions

Complex questions focus on questions that are subject to unpredictable and constantly evolving systems, contexts, and human behaviour, that is, they have the characteristics of complex systems. These 'wicked' questions regularly occur in development that is intrinsically a complex adaptive system (refer to Section 2.2 for a complete discussion of complex adaptive systems and development). Answering complex questions is tricky, and lab-like settings are possible in only a few instances. Typical questions include 'Do women's collectives have the ability to induce the desired transformative change within the system for it to affect women's agency?'" or "Does teaching at the right level crowd-out other interventions to improve school performance?" or "What are enablers and resistors to the idea of a community information exchange to solve the coordination problems in the US healthcare system?" or "Does the conditional cash transfer program have the ability to adapt to changing feedback and shifting contexts over time?"

Answering complex questions typically requires a variety of techniques, and these may not be predetermined in an evaluation (see Section 2.3.1 where we mention a few techniques, including case studies, system mapping, outcome harvesting, process tracing, and qualitative comparative analysis).

While these types of questions naturally arise in the complex domain of the Cynefin framework, the nested nature of the framework means that answering complex questions necessarily involves answering both complicated and simple questions. In other words, complex questions encompass both complicated and simple questions. Thus techniques, such as RCTs and quasi-experiments, can be elements of an evaluation strategy for complex questions.

Unfortunately, what frequently happens is evaluation techniques for complicated questions, are overused to answer complex questions when they only tell one part of the story.

IV. An illustration from COVID-19

Let us illustrate the different kinds of questions outlined above, using the context of vaccine development during the COVID-19 pandemic.

Figure 6: An illustration of the types of questions from the COVID-19 pandemic

	Simple Questions	Complicated Questions	Complex Questions
Vaccine Development	• What is the gene sequence? • What is the reproduction rate, R0? • How were vaccines developed in the past?	• Is the vaccine candidate effective?	• How can we improve upon current vaccine solutions? • How to effectively distribute the vaccine? • How to increase take-up of the vaccine?
Tools/ Methods	• Descriptive Data Analysis • Benchmarking • Secondary or Desk Research • Literature reviews	• Randomized Control Trials (RCT) • Quasi-experiments • Constrained Optimization	?

Evaluation tools for complicated questions are frequently overused in complex questions

Simple questions on vaccine development would include "What is the gene sequence of the coronavirus?" or "What is the reproduction rate of the virus (sometimes called R0)?" or more importantly "How were vaccines developed in the past and what are known innovations in vaccine development?" All these questions either describe the world in the pandemic as it is or the knowledge of vaccine development as we know it. These questions are extremely important, so important that no evaluation that claims to understand the issue of vaccine development can reasonably ignore these simple questions. The gene sequence, reproduction rate, and what has been done, are the foundation for any further analysis. This illustrates that "simple" here is an analytical category concerning our current state of knowledge—of course, a gene sequence is highly complicated in terms of structure, and the historical process of discovery is a very complex one!

Complicated questions around vaccine development would include "Is the vaccine candidate effective?" Again, these questions are crucial. The RCTs by Moderna, Pfizer, and Johnson and Johnson, among others around the world, all were an attempt to answer this complicated question. Needless to say, all these trials were essential. Without knowing which vaccine candidate worked we would not be able to find any resolution to the pandemic.

Lastly, complex questions around vaccine development are many. Consider this complex question—How can we improve current vaccine solutions? This question is complex because it deals with innovation into new areas with unknown responses, and requires the experimentation typically done in research and development. Recall in Section 2.2.2 we showed how innovation has features of a complex adaptive system. Consider another complex question—How can we effectively distribute the vaccine in developing countries? This question is complex because it deals with systems—global and local supply chains that were severely hampered by the pandemic— that are affected by complex geopolitical forces. Finally, another complex question is "How to increase the take-up of the vaccine?" This question is complex because it deals with unpredictable human behaviour and delves into how we can persuade vaccine deniers.

A few observations on complex questions. First, notice that answering any of the questions would involve addressing complicated and simple questions along the way. For example, the complex question on innovations in vaccines would involve testing potential vaccine prototypes to see which is effective (a complicated question) and knowing current ground-breaking innovations in vaccines, such as the mRNA technology (a "simple" question now, again with a history of complexity). The same can be said of the second question on vaccine distribution. This would involve understanding which routes are most effective (a complicated question), and an understanding of how pharmaceutical supply chains currently operate (a simple question). Finally, the third question on vaccine take-up could involve an RCT on estimating the causal effect of public health advertising to persuade people to take the vaccine in a specific context (a complicated question) and a survey on attitudes towards the vaccine (a simple question).

Second, notice that in each case, answering the complicated and simple questions that naturally arise from a complex question is necessary, but not sufficient to resolve the complex question. For example, understanding which shipping routes are effective doesn't resolve the issues around the imbalanced power of developing countries, and the hegemony of vaccine production in developed countries. This harkens back to the discussion in Section 2.3.1 Part I about how understanding parts cannot fully explain the whole. Perhaps more pertinently, it demonstrates the perils of only using techniques of complicated questions to answer complex questions.

Finally, a puzzle that is still unresolved in this section so far is how can we go about answering the complex questions. Moreover, while we cannot only use techniques for complicated questions to address these complex questions, what is the range of techniques we can use, and what process would help us better answer these complex questions?

In the following section, we argue that the answer lies in an Adaptive Evaluation. This will outline both the techniques and the processes that are useful in tackling these 'wicked' questions.

3.2. Adaptive Evaluation

What is an Adaptive Evaluation? So far, we have described aspects of an Adaptive Evaluation, but never in its full form. We can now define an Adaptive Evaluation, outline its core principles, and bring together all the pieces that we partially introduced in previous sections.

An Adaptive Evaluation is a set of processes and techniques designed to answer complex questions in development—questions that are subject to human behaviour, interact with systems and are context-dependent. By complex, we directly reference the complex domain in the Cynefin framework (Section 3.1.1), and within this, how nearly all problems development, innovation, and scaling involve, in essence, complex adaptive systems (see Section 2.1.3).

Thus, an Adaptive Evaluation is based on complexity principles. By that, we mean that an Adaptive Evaluation, takes into account each of the implications of complexity on evaluations listed below (see Section 2.3.1 for a detailed explanation of the implications of complexity of evaluations):

- An Adaptive Evaluation understands that one cannot separately evaluate parts of a system to understand the whole. While an Adaptive Evaluation often aims to understand a specific part of a system (an intervention) this is in the context of how that part interacts with other parts of the system
- An Adaptive Evaluation acknowledges that causality is non-linear and can go in many directions (i.e., displays cyclicality). In Section 3.2.2, we will see that the elements of an Adaptive Evaluation involve many approaches that are designed to tackle non-linear and multi-directional causality
- An Adaptive Evaluation views dynamism as central, is cognizant of the importance of history, and also sees any system as working in a context (i.e., it is nested). In Section 4.2, we describe what we call an Interpretation phase for this purpose.
- An Adaptive Evaluation understands that the timing and nature of impacts may not be known or obvious. In Section 4.2, we will see that the Adaptive Evaluation process is undertaken throughout the life cycle of an intervention, from an idea to innovation to scale. In Section 4.3, we will introduce how we consider an evaluation to be a continuous journey, rather than a measurement at a given point or period in time, and that an evaluation needs to be ongoing in any stage of an intervention.

Note that the core ideology of Adaptive Evaluations is different from standard impact evaluations, for the following reasons:

- Adaptive evaluations ***do not presuppose a hypothesis*** and fix outcomes of interest ex-ante but ***investigate multiple hypotheses on processes and outcomes that emerge from work in the field***. The complex adaptive system that development resides in, necessitates that these hypotheses are generated in consultation with a variety of stakeholders, and can evolve multiple times during the course of the evaluation.
- Adaptive evaluations are ***participatory*** with an ***emphasis on co-creation***. Participation is central to complexity-based evaluations in development. Actors in a complex adaptive system are both key sources of interpretation and knowledge and can be part of the change process itself. Evaluators of an Adaptive Evaluation ***work closely with, or can be embedded within, intervention teams***, and deeply integrated into the process, design, and scaling of the program, working alongside implementers. Objectivity in interpretation has to be sustained by the integrity of the empirical and interpretative process, for example in the transparency of logical and

statistical tests of hypotheses. This is in contrast to evaluators in standard impact evaluations, that at least in principle maintain a distance from the intervention team to maintain independence and objectivity (though in practice there are often quite intimate interactions with intervention teams).

- Adaptive evaluations emphasize a '***learning' approach'***, for both implementation teams and evaluators, rather than a 'testing' approach typically seen in standard impact evaluations. This is an important point of difference. In relatively simple environments, the source of the problem is easy to identify, and it suffices to simply test if a solution works. In complex environments, the source of the problem is unclear or even uncertain, and so one must continuously learn, in an iterative fashion and often in short cycles, and then update to develop an understanding of what works and doesn't through repeated exploration or experimentation.
- Finally, Adaptive Evaluations use a ***variety of techniques*** (both qualitative and quantitative) recognizing that techniques used to address complicated questions may not always apply to complex questions. As vividly observed by psychologist Abraham Maslow, just because you have a hammer, you shouldn't treat everything as if it were a nail. An Adaptive Evaluation is flexible and adapts techniques to the type of question, the nature of the task (innovation or scaling), and the level of complexity.

Figure 7 brings together all the parts of an Adaptive Evaluation. The rest of the section is organized as follows: Section 3.2.1 expands on the different kinds of complex questions addressed in an Adaptive Evaluation and Section 3.2.2 explains the methodological elements of an Adaptive Evaluation.

Figure 7: Previously made statements about an Adaptive Evaluation

What is an adaptive evaluation?

- An adaptive evaluation is a complexity-based approach to systematic learning that is designed to support innovation and scaling change processes in development
- More precisely, it is a set of processes and techniques designed to answer complex questions in development—questions that are subject to human behaviour, interact with systems and are context-dependent

What is the main goal of an adaptive evaluation?

- To understand the complexity of the environment that an intervention or policy inhabits
- To support and inform action (i.e., policy and interventions) that induces meaningful change

Why adaptive?

- It points to a recognition that development, scaling, and innovation all share features of complex adaptive systems (as outlined in Section 2.2)
- It refers to a crucial feature of the evaluation; the emphasis on continuous learning and improvement to *adapt* intervention design to the evolving needs of the end beneficiaries and other stakeholders, and to *adapt* to dynamic contexts and settings as the intervention scales
- The evaluation itself is adaptive, in that the methods and techniques used are intended to adapt to the nature of the task (innovation or scaling), and the complexity of the system
- Adaptive points to two other adaptive elements of the change processes that are outside the scope of this paper but are nonetheless crucial complements to an adaptive evaluation—adaptive leadership and adaptive implementation

What are the contributions of an adaptive evaluation, relative to other evaluations designed for complexity?

- It is more holistic and ambitious because it combines three elements:
    - An evaluation framework for understanding the role of context, history, and theory in outcomes (similar to a realist evaluation)
    - An evaluation framework to support innovation involving rapid measurement, prototyping and experimentation (as in a developmental evaluation)
    - A evaluation framework to support scaling
- It offers a clear strategy, not only of techniques and methods that can be used in each of the three evaluation frameworks above but also of under what conditions is it best to use these techniques and in what sequence.

3.2.1 Types of Complex Questions (addressed by an Adaptive Evaluation)

Adaptive evaluations answer complex questions—questions that are subject to unpredictable human behaviour, systems, and contexts. As noted above, complex questions are contrasted with "simple" and "complicated" questions in the Cynefin framework. This complements the discussion of complexity theory above and provides a useful framing of the different types of evaluation questions of relevance here. The nested nature of the Cynefin framework implies that in practice answering complex questions often involves exploring both complicated and simple questions when specific processes can be separated within an overall system—with implications for techniques that can fruitfully be deployed in an evaluation journey (refer to section 3.1.2. Part III for more details). In this section, we delve into three main kinds of complex questions typically addressed in an Adaptive Evaluation—system diagnostic questions, theory of change questions, and design responsiveness questions. In each area, we are concerned both with describing and interpreting the (part of) the system that is the focus, and the impact and effectiveness of desired (or undesired) changes for intended (and unintended) beneficiaries.

Figure 8: A map of evaluation questions

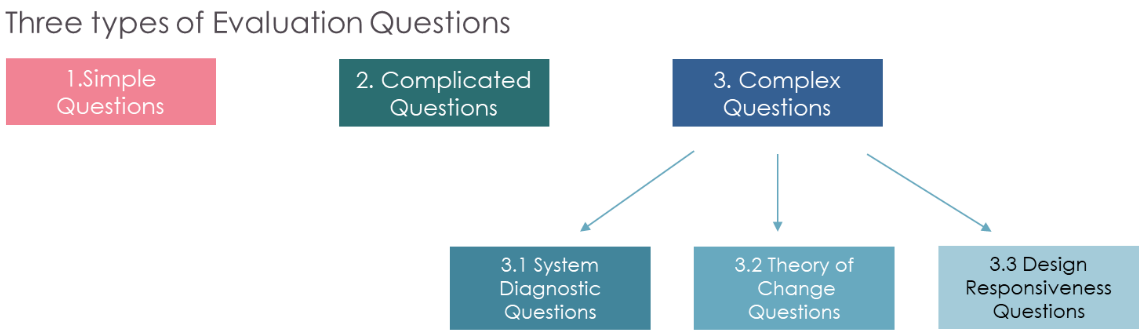


Figure 9: A summary of the three types of complex questions

	1. **System Diagnostic**	2. **Theory of Change**	3. **Design Responsiveness**
Interpretive Questions	• What are the boundaries of the system of interest for the desired change? • Who are the main actors in the system? What are their reciprocal obligations to each other? What are their incentives? • What is the history/trajectory of the system over time? What are the initial conditions of the system? • What are enablers and resistors to the desired change within the system? What potential tailwinds and headwinds outside the system affect the desired change? Is resistance from interest or ideas? • What are the resources, capacity, and capabilities of present in the system to implement the desired change now and in the future? • What potential intervention(s) may lead to the desired change in the system?	• What are some expected causal mechanism(s)/channels for the intervention(s) of interest to achieve the desired change? • What are the explicit/implicit assumptions and theories that explain these expected causal mechanisms? • How are the chosen intervention(s) of interest expected to interact with each other, with existing interventions, and the context of the change?	• What types of feedback is most useful for the intervention(s) of interest? • What structures (s) can be put in place to receive regular feedback, from end beneficiaries and intermediate processes, of the intervention(s) of interest? • What elements, if any, of the design of the intervention (s) can be potentially be modified in response to feedback? • What elements. If any, of the process to implement the intervention can be potentially streamlined in response to the feedback?

Impact and Effectiveness Questions	• Do the intervention (s) of interest have the ability to induce the desired transformative change within the system for it to affect the outcome of interest for the intended beneficiary?
	  • Does the history and the present context support the intervention?
	  • Are incentives and mindsets of the main actors in the system aligned to make the intervention work?
	  • Is the intervention able to leverage enablers of change and control resistors of change? Is the intervention robust to potential threats and tailwinds?
	  • Is the intervention feasible given the resources, capacity, and abilities in the system?
		• How close was the trajectory of change of the intervention(s) of interest to the hypothesized theory of change?
		  • Do the explicit/implicit assumptions or theories that underly the intervention hold?
		  • Which of the expected mechanisms/channels for the intervention(s) worked?
		  • What unintended or unexpected pathways and outcomes occurred?
		• Do the intervention(s) of interest, crowd, out or enhance each other and existing interventions or initiatives?
			• Are the structures to receive regular feedback from end beneficiaries of intervention (s) of interest working effectively? Do they lead to design changes?
			• Have the actors influencing the intervention (s) of interest demonstrated the capacity to adapt/respond to the changing feedback/shifting context over time to improve design and process?

I. System Diagnostic Questions

System diagnostic questions, as the name implies, involve questions about (1) the characteristics and features of the complex adaptive system that the development problem resides in, and (2) the intervention's interactions with the actors, entities, and processes of this complex adaptive system. System diagnostic questions are central to any evaluation with a complexity frame of reference.

System diagnostic questions begin with defining the boundaries of the system of interest (that is, the system in which we want to implement the desired change). In our running example in Section 2.1, the system of interest for improving rural women's empowerment (the desired change) in India was the RDD System (see **Figure 1**). Defining the boundaries of the system of interest involves determining what is internal (endogenous) to the system and what is external (exogenous) to the system. This distinction is made for simplicity and focus—throughout an Adaptive Evaluation, there is an acknowledgement that the system continually interacts with and is affected by actors, entities, and processes outside the system (i.e., the system is open). After defining the boundaries, system diagnostic questions attempt to understand the main actors, entities, and processes in the system, and their reciprocal obligations to one another. This includes discovering tensions and dependencies between various stakeholders. After understanding the main players in the system, system diagnostics questions critically analyse the history of the system of interest and the current conditions under which the system operates. Next, there are questions to identify the enablers and resistors to the desired change inside the systems, and the tailwinds and headwinds that affect the desired change from outside the system. Finally, system diagnostic questions aim to understand the resources, capacity, and capabilities of the system. Answering these questions should give a more comprehensive picture of the system of interest. With this foundation, one can brainstorm interventions that have the potential within this system to create the desired change.

There are also questions over the relationship between system functioning and impacts on core developmental concerns. At one level this is a genuinely complex big question: Is the system effective in fostering human development of the population of concern? Or, does the system as a whole encourage economic dynamism? There are then questions on the effectiveness of particular areas of change, i.e., "interventions" within the system: Is this intervention aligned with system functioning, or will it be rendered ineffective by reactions within the system, or by resource and capacity constraints within the system? Or is there the potential for an area of change, or an intervention, to induce positive (or negative) transformations within the system? These are important questions to address, though in practice it will often make sense to focus on specific pathways within a system, and this takes us to theories of change.

Figure 10: System Diagnostic Questions

Interpretive Questions on the system

- What are the boundaries of the system of interest for the desired change?
- Who are the main actors in the system? What are their reciprocal obligations to each other? What are their incentives?
- What is the history/trajectory of the system over time? What are the initial conditions of the system?
- What are enablers and resistors to the desired change within the system? What potential tailwinds and headwinds outside the system affect the desired change? Is resistance from interest or ideas?
- What are the resources, capacity, and capabilities present in the system to implement the desired change now and in the future?
- What potential intervention(s) may lead to the desired change in the system?

Impact and Effectiveness Questions on the system

- Do the intervention (s) of interest have the ability to induce the desired transformative change within the system for it to affect the outcome of interest for the intended beneficiary?
  - Does the history and the present context support the intervention?
  - Are incentives and mindsets of the main actors in the system aligned to make the intervention work?
  - Is the intervention able to leverage enablers of change and control resistors of change? Is the intervention robust to potential threats and tailwinds?
  - Is the intervention feasible given the resources, capacity, and abilities in the system?

II. Theory of Change Questions

Theory of change questions address how an intervention, or a change process, is expected to lead to changes in other factors, especially final outcomes for the population of concern, within a particular system or context.

Theory of change questions unpack expected causal mechanisms or pathways through which the interventions of interest will achieve the desired change. Once expected causal mechanisms have been identified, questions on theories of change are about understanding the explicit/implicit assumptions and theories (about behaviours, markets, society etc.) that would justify each causal mechanism. Finally, these questions also investigate how the causal mechanisms of different interventions are expected to interact with one another, with existing interventions, and with the context of change.

Impact and effectiveness issues are integral to theory of change questions. These are primarily concerned with verifying and testing the causal mechanisms and their associated assumptions, within the alternative theories. The overarching question is how close is the *actual* trajectory of change to the *hypothesized* theory of change. This question gives rise to other questions, such as (1) which pathways of the intervention were active, and which were inactive, and (2) what were some pathways and outcomes that were unexpected or unintended. Finally, impact and effectiveness theory of change questions include whether the set of interventions enhance or crowd out each other, and how they affect other existing interventions.

Theories of change are, of course, almost always features of evaluations, including what we have called standard impact evaluations. There are two aspects of Adaptive Evaluations that are distinct: the theories seek to take account explicitly of system interactions, or pathways of change within a system; and the methodological design (discussed further below) allows for theoretical testing, updating, and revising within the scope of the evaluation itself.

Figure 11: Theory of Change Questions

Interpretive Questions on theorizing change processes

- What are some expected causal mechanism(s)/channels for the intervention(s) of interest to achieve the desired change?
- What are the explicit/implicit assumptions and theories that explain these expected causal mechanisms?
- How are the chosen intervention(s) of interest expected to interact with each other, with existing interventions, and the context of the change ?

Impact and Effectiveness Questions assesing trajectories of change

- How close was the trajectory of change of the intervention(s) of interest to the hypothesized theory of change?
    - Do the explicit/implicit assumptions or theories that underly the intervention hold?
    - Which of the expected mechanisms/channels for the intervention(s) worked?
- What unintended or unexpected pathways and outcomes occurred?
- Do the intervention(s) of interest, crowd, out or enhance each other and existing interventions or initiatives?

III. Design Responsiveness Questions

Design responsiveness questions are concerned with the ability of the implementing actors to regularly receive and collect useful and actionable feedback, and importantly, act on the feedback to adapt designs to this feedback. These questions are concerned with modifications that can improve the effectiveness of an intervention. They also engage with the more ambitious institutional task of building the mechanisms and operating processes required to sustain an organizational culture of continuous learning.

Design responsiveness questions involve understanding how feedback works within the institutional system as it is now—that could range from an organization that supports formal design sprints, to one that is driven by standardized rules, but is nevertheless concerned with actual impacts. Then the questions of effectiveness are essentially the same as those within a theory of change but now are embedded with a process of assessing and modifying design choices over time, and with how this relates to system functioning.

Figure 12: Design Responsiveness Questions

Interpretive Questions on design responsiveness

- What types of feedback is most useful for the intervention(s) of interest?
- What structures(s) and processes can be put in place to receive regular feedback, from end beneficiaries and intermediate processes, of the intervention(s) of interest?
- What elements, if any, of the design of the intervention (s) can be potentially be modified in response to feedback?
- What elements. If any, of the process to implement the intervention can be potentially streamlined in response to the feedback?

Impact and Effectiveness Questions on design responsiveness

- Are the structures to receive regular feedback from end beneficiaries of intervention (s) of interest working effectively? Do they lead to design changes?
- Have the actors influencing the intervention (s) of interest demonstrated the capacity to adapt/respond to the changing feedback/shifting context over time to improve design and process?

IV. Discussion of the Types of Complex Questions

System diagnostic, theory of change, and design responsiveness questions capture the categories of issues within complex questions. Each type directly addresses some aspect of complexity. System-diagnostic questions unravel the complex nature of systems, exploring the tendencies of the system, its evolution over time, and how the system may react to the introduction of a novel intervention, giving us a sense of whether an intervention would disrupt or perpetuate the status quo. Theory of change questions are structured around pathways of change within the system, typically associated with one part, or one starting point for an intervention. They grapple with the many assumptions and theories that seek to explain a causal mechanism and explore the numerous pathways through which an intervention could affect the desired change. Lastly, design responsiveness questions address a feature that is at the heart of complexity—adaptation. Design responsiveness questions systematically link assessment of effectiveness to processes of modification of designs, in response to feedback, and may engage with the broader issue of understanding, and supporting, learning processes within the system.

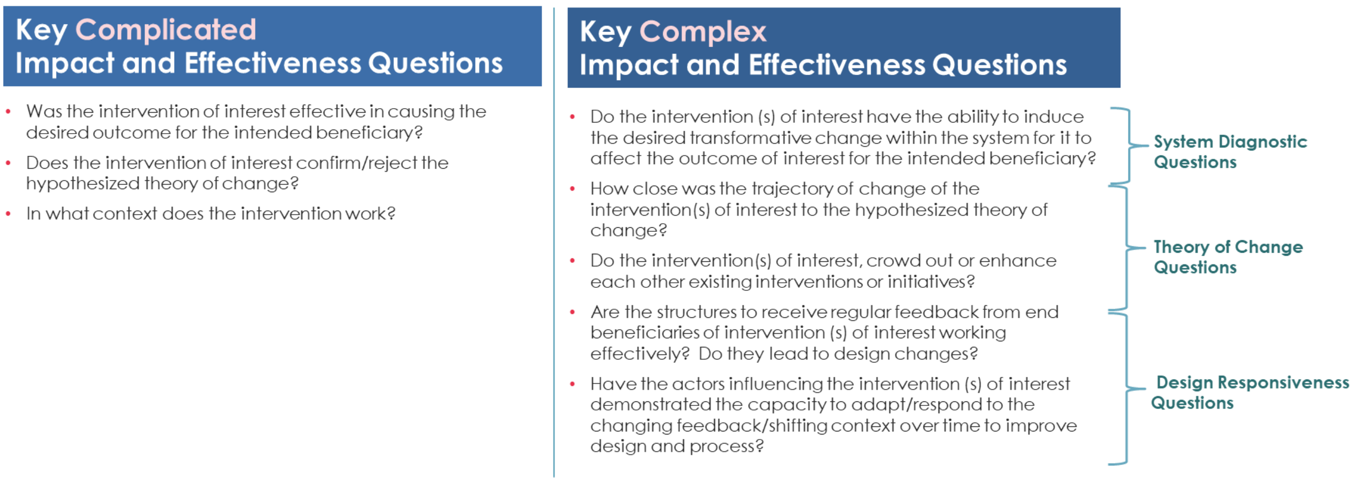
Finally, a careful examination of all the questions in **Figure 9** together reveals how complex questions across the different types respect the implications of complexity on evaluations highlighted in Section 2.3.1. Complex questions not only include questions about the intervention (a part of the system) but also about the intervention's role in the system (the interaction between the parts of the system). The complex questions in **Figure 9** include a question about the trajectory of the system and initial conditions, therefore honouring the role of history in creating a credible account of transformation. Moreover, the complex questions in **Figure 9** consider multi-directional and non-linear causality throughout, imperative in any evaluation with a complexity frame of reference. Last but not least, complex questions are dynamic, and so implicitly understand that the nature and timing of impacts may be unknown or uncertain.

V. Discussion on Impact and Effectiveness

"Development Effectiveness" has always been central to the international development space. The "empirical (or credibility) revolution" in economics—that came to the forefront in the 1990s—brought a hugely important emphasis on the careful empirical analysis of policy and intervention design. However, in the past 20-30 years this has become associated with specific interpretations and techniques, especially in the core development industry. Funders and donor agencies often require an assessment of impact with specific techniques *before* giving grants to governments and social enterprises.

Typical evaluations (commonly referred to as simply, "Impact Evaluations"), have used RCTs where feasible, or other quasi-experimental methods with strong identification strategies. As one element of the broader "empirical revolution," these have had an important positive impact on a range of specific areas. However, this approach has appropriated the term "impact", both amongst practitioners of the method (where it makes sense in the context of the statistical prism of the broader approach), and amongst much of the funding and donor community to the point, where many assume that any question of impact and effectiveness must involve an RCT. As we have argued here, RCTs are excellent instruments for simple and complicated problems, but not for complex ones, where they have at least to be combined with other methods. **Figure 13** below illustrates the difference between complicated impact and effectiveness questions typically addressed in a standard impact evaluation and the complex impact and effectiveness questions typically addressed in an Adaptive Evaluation.

Figure 13: Differences between Complicated and Complex Impact and Effectiveness Questions



There are quite a few points of difference between complicated and complex impact and effectiveness questions. First, complicated questions on impact and effectiveness are generally assessments for one (or sometimes two) periods, occasionally assessing longer-term changes within the same testing framework. By contrast, the assessment of the impact and effectiveness of complex questions is *dynamic,* both in the sense of considering multiple periods and more importantly in allowing for shifting frames of causal interaction. Second, complicated impact and effectiveness questions focus on the intervention in isolation, by design. On the other hand, complex impact and effectiveness questions place the emphasis not only on the intervention but also on how the intervention interacts with the system, how one can navigate change within the system and sometimes how the system itself can change or transform through the process. Third, complicated questions on impact are focused on *testing* an intervention along with its associated theory. On the contrary, while complex questions on impact also involve testing an intervention along with its associated theory, the primary focus is on learning and the ability to adapt. Finally, complicated impact questions typically consider one intervention (or a few, each with different designs), while complex impact questions consider many interventions, often changing over time.

3.2.2. Methodological Elements of an Adaptive Evaluation

In the previous section, we discussed the different types of complex questions typically addressed in an Adaptive Evaluation: how does the system function? What are the causal channels of change within the system? And how do actors respond to feedback to modify designs? In this section, we provide an outline of methodologies for each of these questions. An Adaptive Evaluation does not involve one technique, but rather an analytical approach within which alternative methodological techniques can be applied and matched to the specific problem, context and potential data.

Figure 14 summarizes both the overall approach and the menu of techniques for the three elements of an Adaptive Evaluation. In almost all cases an Adaptive Evaluation will use mixed methods, both in the general sense of applying more than one of the menus and in the specific sense of combining quantitative and qualitative methods. Qualitative methods are more amenable to an open-ended exploratory assessment within a complex, evolving system. They are particularly powerful at outlining the breadth of possibilities, exploring potential mechanisms in change processes, and providing a rich, nuanced understanding of how the inextricably linked social, historical, political, and economic contexts may have contributed to outcomes. Quantitative methods are useful in exploring and identifying which of the breadth of possibilities identified in a qualitative study are most important in explaining why and how the intervention succeeded. Quantitative methods naturally benefit from larger sample sizes, and therefore are more generalizable, in the specific statistical sense, that much more can be inferred about the population of interest from the sample. Mixed methods allow researchers to triangulate results, that is, to confirm (or negate) qualitative results with quantitative data, and vice-versa.

Figure 14: The Methodological Elements of an Adaptive Evaluation

	Description	Question Type	Illustrative Techniques
Systems Based Approaches	• Maps the history of the system and initial conditions • Make explicit interrelationships between key actors, entities and processes • Identify enablers and resistors of change • Evaluate the resources and capacity of the system	• System Diagnostic	• Actor Mapping • Circular Interviews • Roleplaying • System Dynamics • Agent-Based Modeling • Network Analysis • Case Studies
Theory Based Approaches	• Unpack mechanisms in-between the intervention to the outcome • Extract and test theories and assumptions behind the causal mechanism • Allow engaging with several interventions, hypotheses, and their interactions	• Theory of Change	• Logic models (fishbone and Ishikawa diagrams) • (Bayesian) Process Tracing • RCT/Quasi-Experimental Designs
Iterative Approaches	• Involves rapid prototyping, experimentation, and measurement in short sprints • Enables testing multiple designs and actively learning to improve designs • Requires building systems for learning and regular feedback	• Responsiveness	• A/B Testing • Beneficiary Interviews • User Journey • Needs Assessment • Process Mapping/Monitoring

We use an overall mixed-method approach in an Adaptive Evaluation for two main reasons. First, mixed methods are particularly suited for complex environments. Complexity theory makes it abundantly clear that no one dataset, methodology, or approach, is sufficient to provide a panoramic view of the system (recall the critique of economic models in Section 2.3.1 Part I, which emphasizes how part of a system cannot explain a whole) [Walton, 2014]. Second, a *mixed* approach is crucial to the philosophical underpinnings of an Adaptive Evaluation. As described at the very beginning of Section 3, an Adaptive Evaluation is open-minded with techniques, recognizing that techniques to answer complicated questions may not always work in answering complex questions. Embedded in this philosophy, is a preference for methodological flexibility and eclecticism, which the *mixed* nature of mixed methods aptly encapsulates.

We now delve into some core methods for the three elements of an Adaptive Evaluation—system diagnosis, theory-based interpretation of change, and design responsiveness. This isn't intended to summarize all methods in **Figure 14**, but rather to sketch some of the main methodological approaches.

I. Methods for System Diagnosis

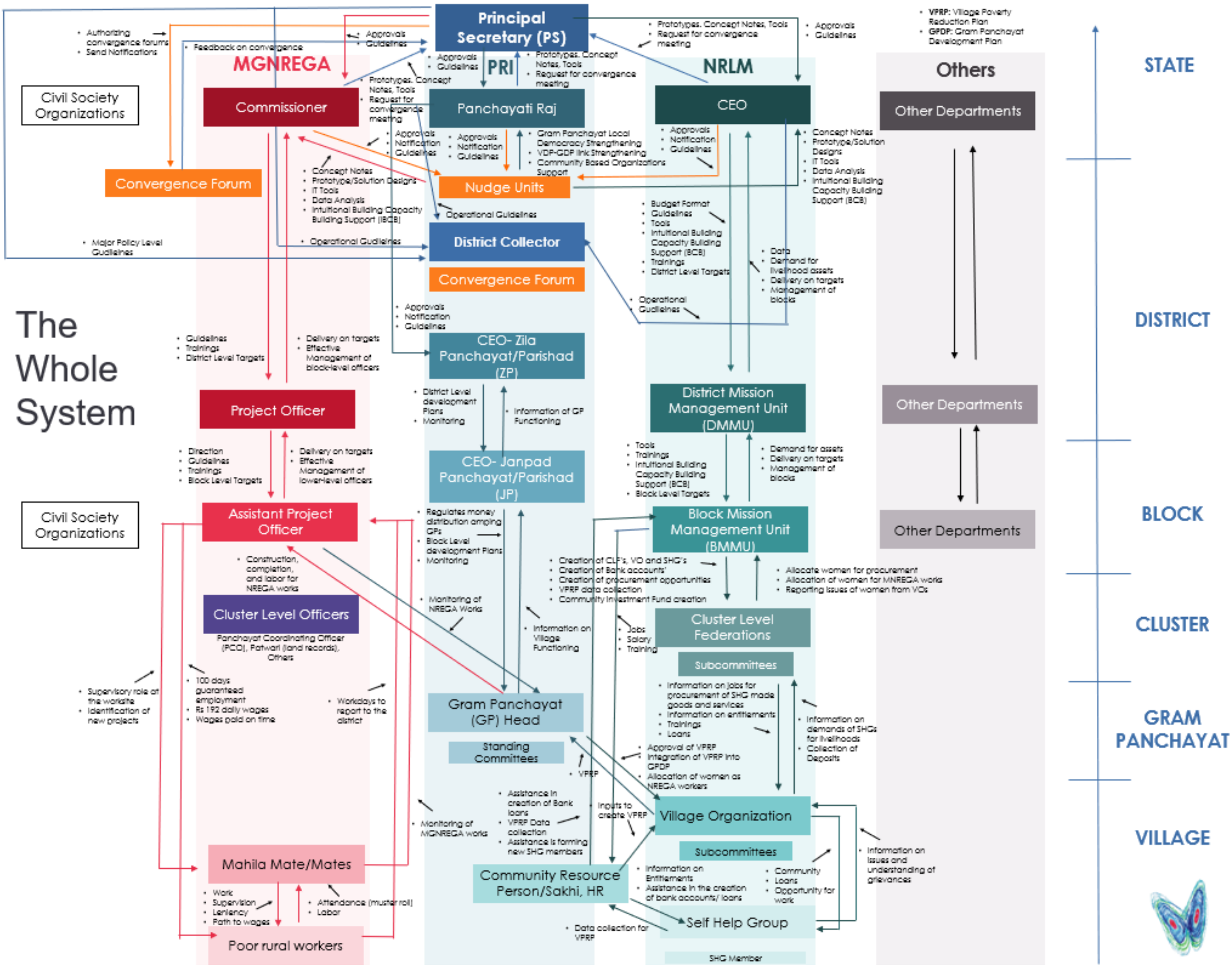
System-based approaches broadly involve techniques to (1) understand the characteristics, dynamics, and functioning of the system of interest, and (2) understand how an intervention, or change process, interacts with the system of interest. Interpreting a system, which addresses the first part, involves defining the boundaries of the system of interest, understanding the key actors and entities in the system (including their reciprocal relationships, tensions, and dependencies), documenting the history of the system, determining the enablers and resistors of change, identifying the tailwinds and headwinds that affect the system from the outside, and figuring out the resources and capacity of the system.

System diagnosis can involve a range of techniques, with varying degrees of formality, up to fully developed, mathematical system dynamic models. Moving to greater formality comes with the trade-off of a loss of transparency, and a lack of flexibility—and can actually be inconsistent with the emergent properties of a system. For practical evaluations, we recommend combining participatory and qualitative system diagnosis, with an assessment of insights from the relevant theoretical and empirical literature on system behaviour. Here we expand on three techniques for qualitative system approaches—participatory system actor mapping, role-playing, and circular interviews (these draw extensively on design thinking.)

An effective technique for initiating a system diagnosis is a 1–2-day **participatory system actor workshop**. A typical participatory system actor workshop involves gathering actors closely involved with the problem, intervention, or desired change in one room, with the initial task to (1) list and describe all the relevant actors and entities and their roles (**stakeholder mapping**), and (2) map the reciprocal obligations of all the actors and entities with each other, starting with those of greatest interest, including the final beneficiaries for an actual or potential intervention (**actor and relationship mapping**). The success of this technique depends on convening 10-15 system actors, or participants, who have as diverse a lens on the problem, intervention, or desired change, as possible. Ideally, they are actors with extensive experience in different parts of the system.

Participants describe the motivations and incentives for the relevant actors and entities and work through inter-relationships until a map of the system appears (facilitation certainly helps, especially in the initial stages). This works especially well in practical design mode, with toy figures and plenty of post-its and markers. It can then be captured digitally. The cover page of this paper is of a participatory system actor workshop in action. **Figure 15** shows a digitized version that is based on the Rural Development Department (RDD) system introduced in Section 2. An example of an actor map that was made from the inputs of a participatory system actor workshop is below (see **Figure 15**).

Figure 15: An example of an output from a participatory system actor workshop conducted by Imago Global Grassroots. This is a system description of the Rural Development Department (RDD) in Madhya Pradesh, India

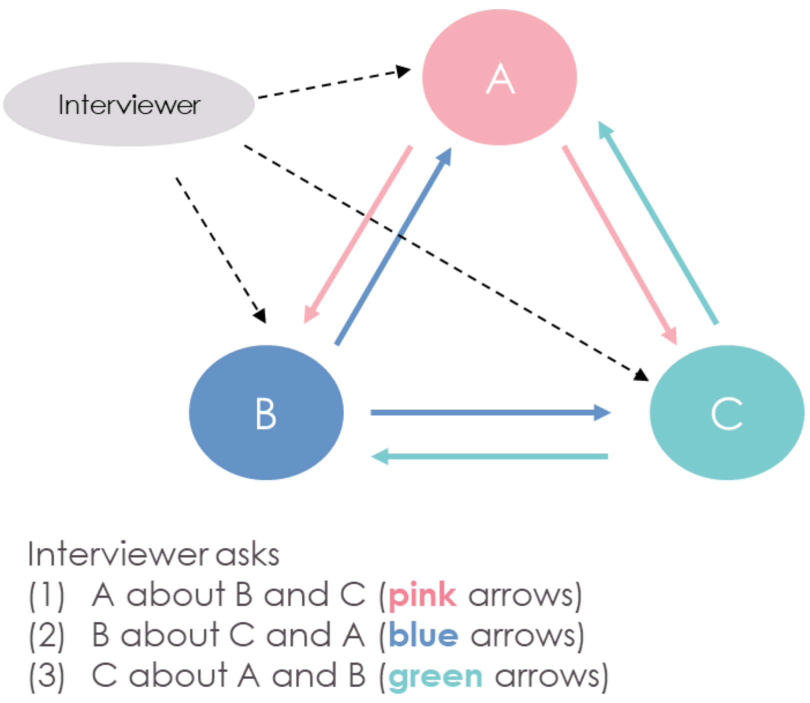


The next task in a participatory system actor workshop is to critically examine, as a group, the completed actor map created together. At this point, the actor map has a lot of the reciprocal obligations mapped in all their complexity. Now one can discuss which parts of the system are functioning as expected, and which are malfunctioning (these are often called blockages). It is then useful to assess actual or potential interventions (or levers for a change). This is linked to an assessment of which actors/entities are enablers of the desired change and which are resistors. One can also include here a discussion of forces outside the mapped system that may affect the desired change (as discussed in Section 2.1.2 Part I and Part II, complex systems are open and nested in other systems).

A direct complement to system mapping is **roleplaying**, in which different actors associated with the problem, intervention, or desired change, are brought together in a room, and randomly assigned roles and a specific blockage to act out. Once the roleplay is concluded, participants pause to reflect on what they heard and saw. Roleplaying is a powerful technique because it builds empathy for the different actors involved (empathy that can then feed into the definition of the problem and design of the intervention). Moreover, roleplays can shed light on the complexity of the problem at hand, and the myriad forces at play.

This participatory work can give evaluators insight into which parts of the system to zoom into for a more in-depth understanding of system functioning. **Circular interviews** can be conducted with a smaller subset of actors of the system to understand their reciprocal obligations to each other, and also delve deeper into the tensions and dependencies between them (Hummelbrunner, 2000). As an example, consider the education system, and three key actors within it—parents (A), teachers (B), and the municipality education department (C). A circular interview would involve interviews asking the parents (A) about their relationship and/or perceptions of the teachers (B) and the municipality education department (C). Similarly, the interviewer would ask the teachers (B) about their relationship and/or views about parents (A) and the education department (C). Finally, one would interview the education department (C) and inquire about the parents (A) and the teachers (B). A triangulation of each actor's viewpoint of others helps expound not only their mutual obligations but also their differing perspectives on the issue at hand. Insights from these circular interviews can then be used to update the actor map from the participatory system actor workshop

Figure 16: Circular interviews (Hummelbrunner, 2000)



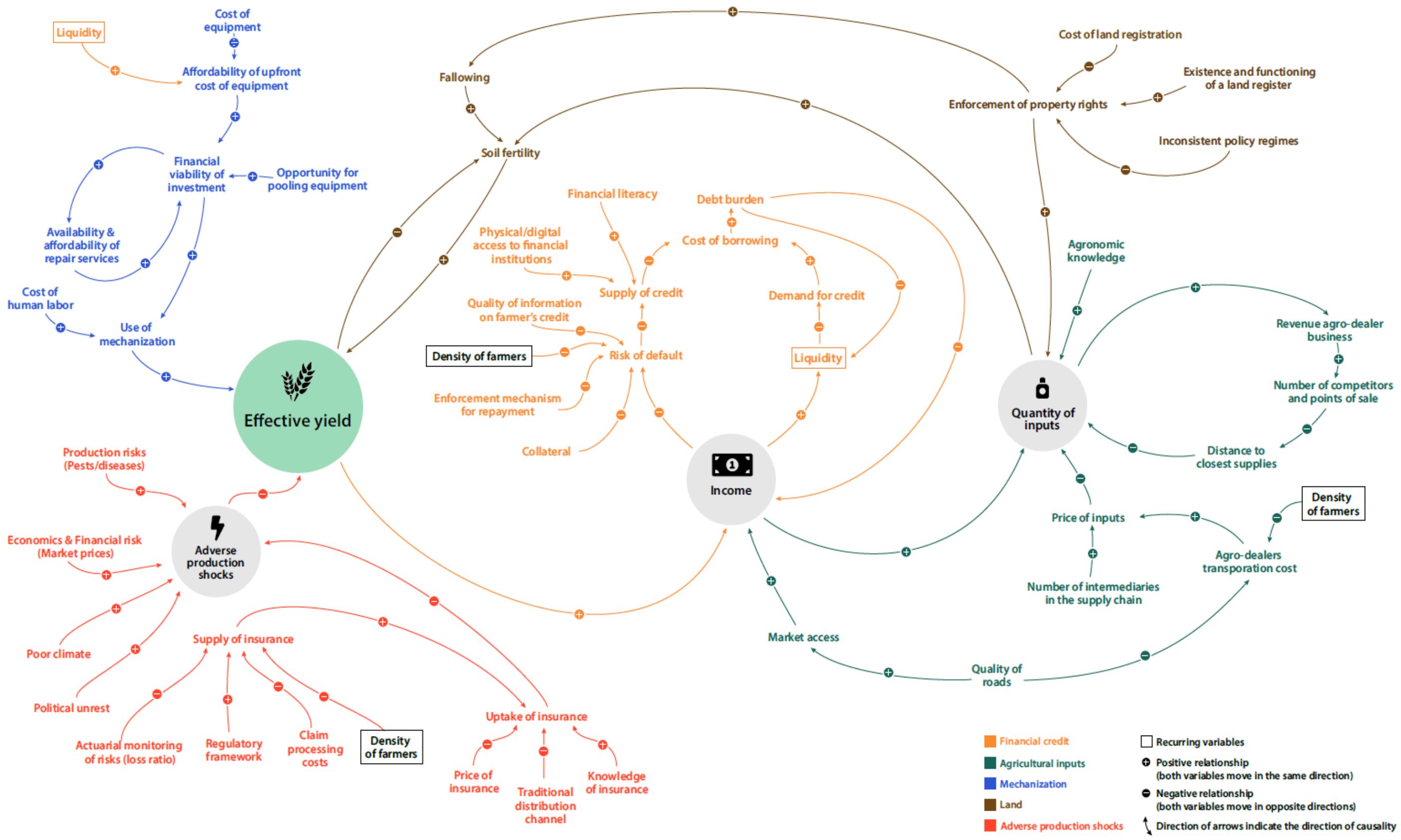
The strengths of participatory system actor interviews, roleplay and circular interviews lie in providing a grounded understanding of the system from the people closest to the problem—the stakeholders. Together, these techniques surface actors' obligations and incentives, and their tensions and dependencies. These can be combined with theory and empirical literature relevant to system behaviour. For example, in the system described in **Figure 15**, a central part concerns the multiple layers of the Indian state. There's a large literature on how the Indian state works! This is embedded in the broader body of literature on state functioning from many disciplines, especially from political science, public administration and economics.

It would make no sense to even attempt a summary of these strands of literature here. Rather the point is to link fieldwork to existing work and integrate interpretations. For example, a recent study of part of the same system of rural development in **Figure 15** combined qualitative interviews and focus groups of all the main actors in the local system, with a reading of both governmental texts and existing literature on societal and state functioning (Bhanjdeo et al., 2021). A central interpretative finding in this study concerned the inherent tensions between the hierarchical features of the Indian state (a form of hierarchical principal-agent relations), associated target culture, and the participatory goals and notional practices of the program.

The work of linking to literature also involves understanding the history of a system. This can be done by reading and analysing historical accounts and studies that already exist. Purposive interviews with those actors that have been in a relevant organization or entity for a long time or used to formerly be in important influencing positions in the system, can also help trace the history of the system. Last but not the least, initial conditions and history can be described using **ethnographic studies** that provide a very detailed account of the norms, behaviours, ways of being, and ways of working of important actors in the system. Ethnography is typically conducted in reference to a particular context, making it a useful tool to understand initial conditions.

We finally illustrate a further technique of system dynamics, that seeks to formalize the positive and negative feedback loops in the system. System dynamic techniques are natural extensions of a system mapping exercise and can be developed as a conceptual illustration of processes or can have quantified parameters. **Figure 17** below has an example of a qualitative system dynamic map that attempts to understand the key determinants of low agricultural productivity. The overall flows in the system are the following. High costs of borrowing increase the debt burden, thereby reducing the income of already cash-constrained farmers. This low-income generating capacity results in poor investments in important inputs (such as fertilizers, pesticides etc.) and mechanization, both of which negatively impact agricultural productivity. Finally, adverse production shocks due to disease or inclement weather and lack of insurance, coupled with insecure land rights, further constrain productivity. System Dynamics can also quantitatively model these relationships and feedback loops, to understand how these variables would move together with time and understand the net result of the many positive and negative feedback loops identified.

Figure 17: A system dynamic map of agriculture productivity (El Houda et al., 2020)



In this section, we have described a variety of system-based approaches to answer interpretive system diagnostic questions, including, participatory system actor workshops, circular interviews, roleplaying, system dynamics, and case studies. This is by no means an exhaustive list of system-based approaches, but rather a collection of tools, we have found to be useful in answering -system diagnostic questions.

II. Theory-Based Interpretation of Change Processes

The next, rather fundamental, step in the evaluation is understanding change, and especially—in a policy context—the potential causal effect of "interventions" and associated processes on outcomes of concern. While we have emphasized the centrality of understanding the system, it is almost never the case that the evaluation is seeking to assess overall system dynamics. To do this would be a hard to impossible task for a complex adaptive system, for all the reasons discussed above. Rather the philosophy of this part of the evaluation is to have a structured analysis of change processes, especially those induced by intentional action in parts of the system (e.g., by a policy-maker, bureaucrat, civil society actor, or entrepreneur within the system). This necessarily involves some selection of a part of the system and abstraction from the overall system behaviour, whilst designing an evaluation strategy that is open to both influences from system functioning, and potential induced changes within the broader system.

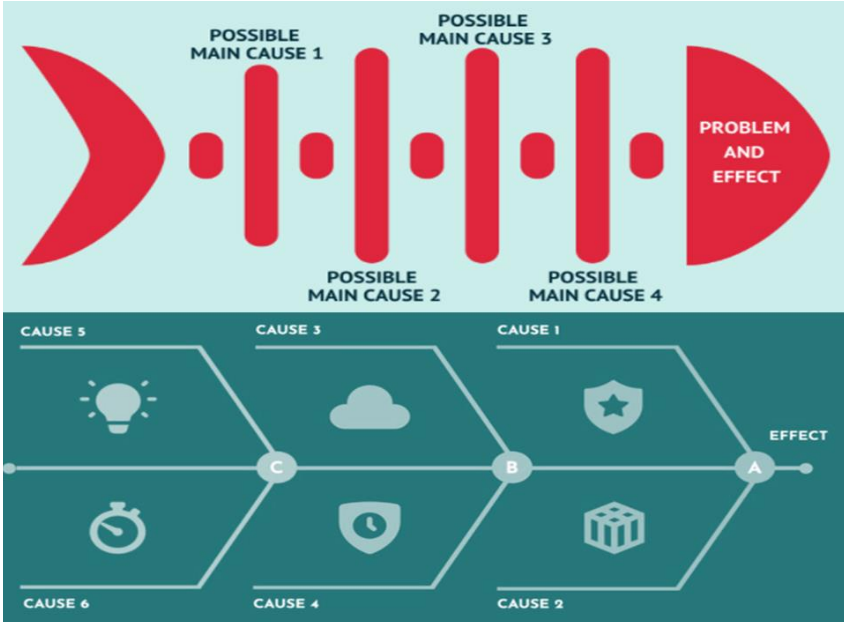
In this section, we argue that a core design element of an Adaptive Evaluation is the development of a logic model, based on relevant theories, that can be assessed by a variety of statistical and logical tests (ideally using mixed quantitative and qualitative methods), and can potentially be updated as information comes in over the course of the evaluation. This is often referred to as a theory-based evaluation.[18] Theory-based approaches are particularly useful in evaluations based on a complexity frame of reference because they consider the intervention in the wider socio-economic-cultural context and also incorporate the actions of other interventions tackling the same issues. Even though theory-based approaches focus on an intervention (i.e., a part of the system), they do not treat this as formally separable, i.e., they are not oblivious of how the intervention interacts with and resides with its wider environment (i.e., the system as a whole). These typically have two parts—*conceptual* and *empirical*. The *conceptual* part entails an explication of the theory that underpins an intervention (Coryn et al.,2011). The *empirical* part entails collecting evidence to test (validate or invalidate) the hypotheses in the logic model.

It is possible to jump straight to an intervention of interest (a new pedagogy in schools, a credit programme etc) and build the logic model around this, with some attention to system interactions. However, we recommend two prior steps involving (1) causal mapping, and (2) assessing pathways through the system.

Causal mapping is a theoretically and empirically informed assessment of the range of influences on an outcome of interests. Take the goal of expanding women's agency. There are a range of alternative channels of influence on this, including household wealth, power relations within a household, links between this and patriarchal culture (in the village and broader society), the rise in women role models, cash transfers and employment opportunities for women, the dynamics of women's groups in the formation of critical consciousness, and more. An **Ishikawa or fishbone diagram** is a useful visual tool for mapping the variety of causal pathways. (see **Figure 18**). If there is an overarching theory of social and economic behaviour, this can take the form of a structured "diagnostic tree", for example categorizing influences on women's labour force participation between supply and demand effects. The final step would be to think about assumptions that must be true to make progress toward agency (e.g., a willingness among men to allow women to express their freedoms, and a political will to discuss women's issues).

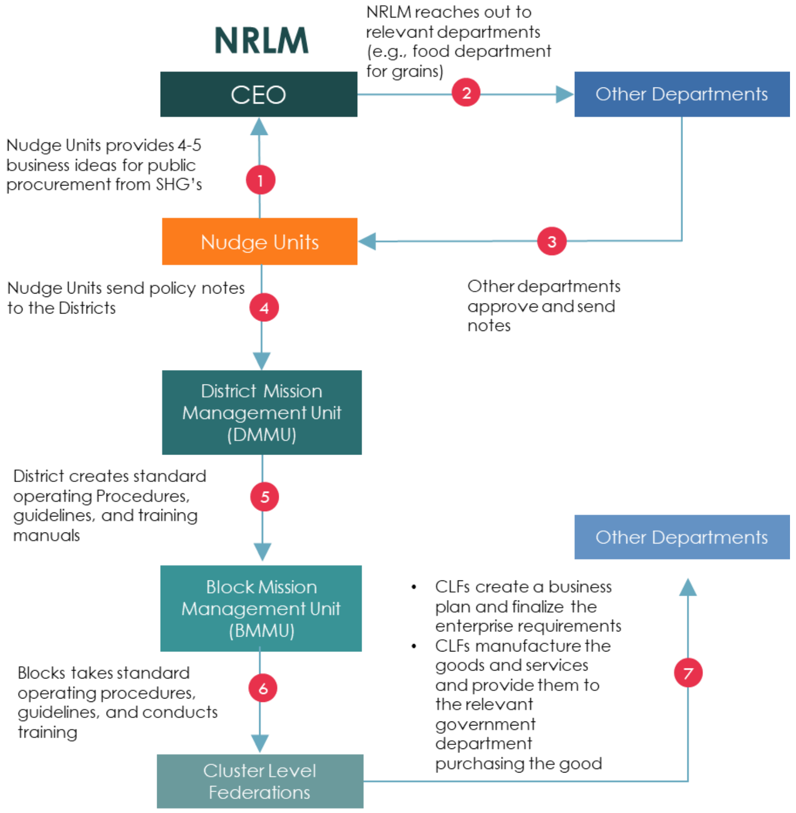
[18] There are three main kinds of theory-oriented approaches—*theory-driven* evaluations developed by Chen and Rossi, *theory-based* evaluations developed by Carol Weiss, and realistic evaluation, developed by Pawson and Tilly. The share most of the tenets of theory-oriented approaches, including and understanding of whether an intervention caused the outcome, but also why and how, and the emphasis of considering the intervention within a context (Stame, 2004). The differences is in the role assigned to theory. In the words of Stame— "*For Chen and Rossi good theories should substitute for no theory; for Weiss, better theories should substitute bad ones; for Pawson and Tilley, theories become good thanks to what actors do about them*."

Figure 18: Fishbone, or Ishikawa, Diagrams



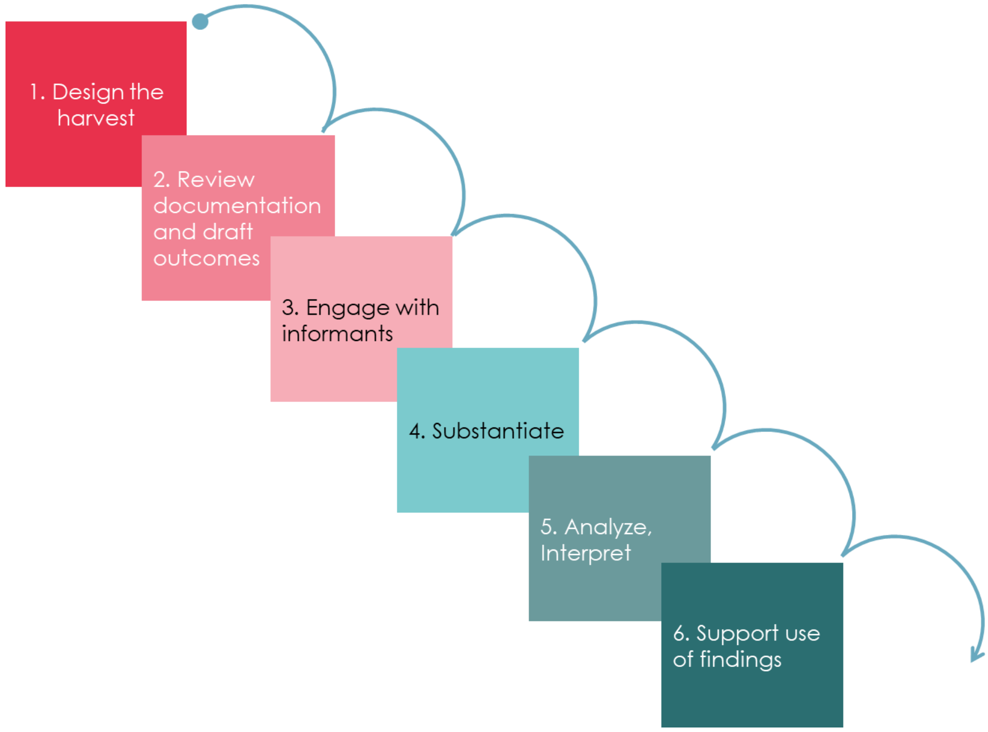
Now this provides a context for focusing on one or more causal sequence that is of practical policy relevance. However, this has to be implemented within the system. How will the system respond? How can intentional actors navigate the system, or even change system functioning? The system diagnosis can support this in several ways. The system map is one point of departure, now with a more specific focus on the intervention, that allows exploration of **navigational pathways**, and interactions within the system. Mapping the flow of action through the system helps identify sources of leverage and resistance (See **Figure 19** for an example of the flow of action of an intervention to increase public procurement of goods and services produced by self-help groups as a means of income generation). This can then be taken to traditional stakeholder analysis of support and resistance and, for example, interpretation of decision-making behaviour from existing literature or fieldwork. It has to be linked to an understanding of the resources and capacity available in the system.

Figure 19: The flow of action of an intervention to increase public procurement of goods and services produced by self-help groups in Madhya Pradesh



An intervention—an intentional design change—always has some context, and often a history of relevant attempts to effect change. For retrospective analysis of both causal processes and patterns of support and resistance within a system, a useful participatory technique is **outcome harvesting**. Instead of starting with a theory of change and then assessing this against evidence, outcome harvesting reverses the process. An outcome harvest begins with a group of relevant actors putting together outcome descriptions that are parsed from reports, previous evaluations, or field experience. These outcomes are substantiated, working backwards, to determine the underlying theories of change, that is how and why the intervention led to the outcome (Wilson-Grau & Britt, 2012). The strength of an outcome harvest is that it thrives in complex settings, especially those in which the aim of an intervention and its actions are not concretely defined. Its participatory nature ensures that it builds evidence from the viewpoints of diverse stakeholders. Outcome harvesting is particularly valuable in finding the unintended consequences of an intervention. **Figure 20** outlines the 6 steps of outcome harvesting

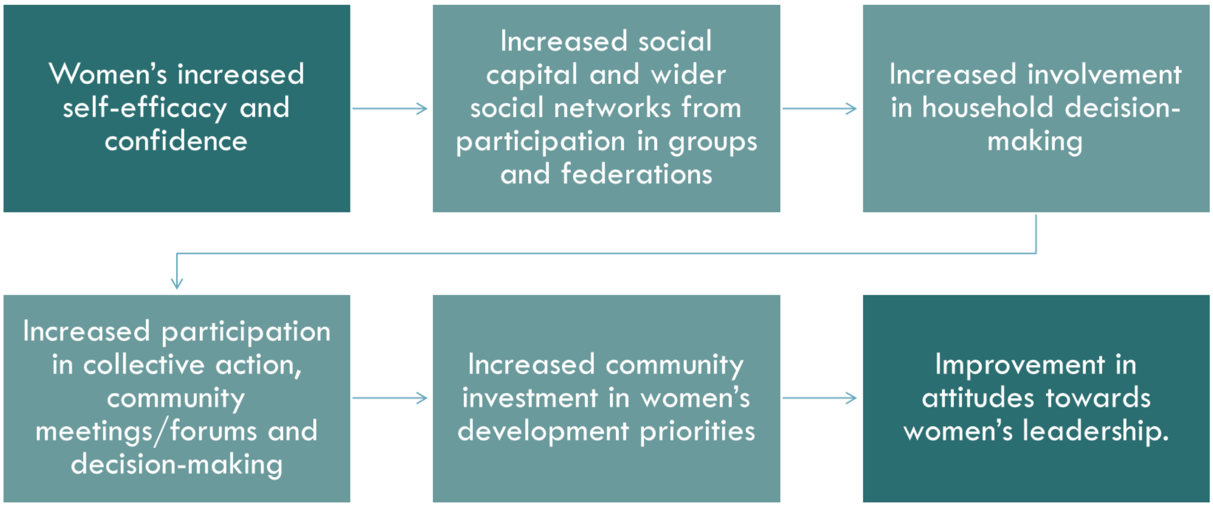
Figure 20: The six steps of Outcome Harvesting



Now let's return to the core methodological instrument, a **logic model**. This will typically involve a focus on one or two causal pathways that are the specific intention of the implementing actors and relate to one or two pathways in the causal map/Ishikawa diagram and associated navigational pathways through the system. In representation, a logic model is typically abstracted as a linearized sequence. The logic model helps understand (1) expected casual mechanisms/channels for the intervention to reach the desired change, (2) explicit implications/assumptions and theories that explain the causal mechanism, (3) the interactions of interventions of interest with other interventions, and (4) assumptions and preconditions that are necessary for progress towards desired objectives. Again, this sequence is ideally developed from a mix of literature review, quantitative, participatory, and other qualitative techniques involving the implementation team and key stakeholders.

Let us consider a programme to improve women's agency. This first step would be to outline the core overarching theory and assumptions about women's agency (e.g., theories about gender norms, patriarchy, and household behaviour concerning men and women). The second step would be to build theories of change around specific interventions, say, women's collective groups, that is, explicate the causal chain that explains how and why women's collective groups lead to women's agency (see **Figure 21**).

Figure 21: An example of a Theory of Change for Women Collective Groups, illustrated in a logic model



Logic models form the structure for empirical assessment, or, more broadly, empirical learning from the processes of change induced by the intentional change (i.e., the "intervention") This involves taking the theory of change to testable hypotheses against evidence. This requires answering not only what worked and didn't in an intervention, but also, why the intervention was able to achieve the desired impact (or not), and how did the intervention achieve those desired impacts (or not) [Carter, 2012]. In other words, answering what worked, involves analysing *whether* the intervention worked (did the intervention 'cause' the desired change in outcomes), but also *what was it about the intervention* that made it work.

There are a variety of techniques that can be used to test the theories, assumptions, and hypotheses of a causal chain or logic model. **Process tracing** is a core technique that provides an overall structure for empirical testing. This is typically used with specific hypotheses, implied by each step of the logic model, against unfolding evidence (generally changes in conditions over time), that constitute varying degrees of logical test to support or reject the initial, or alternative hypotheses. This can be integrated with the randomized allocation of treatments across groups (as in **RCTs**) either for a specific step, or to assess the statistical relationship between the initial intervention and a final outcome.

Note that we have a different emphasis relative to the approach in a standard "impact evaluation" that has been privileged by economic and other evaluations in development policy work in the past two to three decades. This tends to see an RCT as ***the*** instrument, or "gold standard" for evaluation—with better RCTs being accompanied by careful process monitoring to support interpretation. In contrast, we see randomization as one option amongst various potential tests for part of a theory of change, that provides useful information if it can be included, precisely because of its benefits in statistical attribution for that part of the system. In the absence of the conditions for an RCT, however, we remain open-minded to other pieces of evidence and update accordingly. Perhaps most important, the evaluation questions to be tackled should not be limited to what can be randomized or found in natural-quasi experiments! The limitation of an exclusively RCT type approach in the context of an adaptive system is that it is a weak instrument for exploring multiple hypotheses and multiple steps within an evolving system. Moreover, it is often not an option, practically, to randomize treatments.

There are plenty of practical manuals for process tracing and RCTs (Bennett & Checkel, 2015; Collier, 2011; Glennerster & Takavarasha, 2013). We expand briefly on process tracing here, as a technique that is better known in political science and other non-economic social sciences than core economic development work. Process tracing provides a framework of logic tests to parse what any piece of evidence can say, not only about the main hypothesis but also about alternative hypotheses (Collier, 2011). There are four types of process tracing tests based on whether the evidence is necessary or sufficient for affirming causal interference (see **Figure 22** below). Bayesian Process Tracing (BPT) uses Bayesian logic to update initial priors about a particular hypothesis in light of the strength of new evidence collected (Raimondo, 2020). Process tracing allows us to assess *any* type of evidence and what it implies for the main and alternative hypotheses. Moreover, process tracing can be more effectively used to test the entire causal chain outlined in the theory of change. As with any careful technique, it does require a lot of discipline to undertake and a close relationship with an implementing agency.

Figure 22: The four Process Tracing Tests

		Sufficient for Affirming Casual Inference	
		No	**Yes**
Necessary for Affirming Casual Inference	**No**	**1. Straw-in-the-Wind** **a. Passing**: Affirms relevance of hypothesis but does not confirm it. **b. Failing**: Hypothesis is not eliminated but is slightly weakened **c. Implications for rival hypothesis** **Passing**: *slightly* weakens them **Failing**: *slightly* strengthens them	**2. Smoking Gun** **a. Passing**: Confirms hypothesis **b. Failing**: Hypothesis is not eliminated but is somewhat weakened **c. Implications for rival hypothesis** **Passing**: *substantially* weakens them **Failing**: *somewhat* strengthens them
	Yes	**3. Hoop** **a. Passing**: Affirms relevance of hypothesis but does not confirm it. **b. Failing**: Eliminates hypothesis **c. Implications for rival hypothesis** **Passing**: *somewhat* weakens them **Failing**: *somewhat* strengthens them	**4. Doubly Decisive** **a. Passing**: Confirms hypothesis and eliminates others. **b. Failing**: Eliminates hypothesis **c. Implications for rival hypothesis** **Passing**: *eliminates* them **Failing**: *substantially* strengthens

A complementary, technique we would highlight in this section is an analysis of **positive deviance.** This is based on observational analysis of patterns of change, where there is substantial variation across implementing organizations (e.g., schools, health centres, police stations), or geographic areas. "Positive deviance" typically uses standard statistical techniques to identify positive (or negative outliers), preferably conditional on other observable conditions, such as local economic conditions, household resources and inputs in implementation. Then other techniques can be used to explore what works and why, for the positive outliers—for example using interviews, outcome harvesting and other elements of retrospective process tracing. To the extent that there are identifiable features of success, these can then be extended to other parts of the system, with then the further observation of effects over time, to test hypotheses over the replicability of processes in the positive outliers to other contexts.

We would finally return to the overall philosophy of ***learning within a dynamic system***. Some of the above techniques—logical tests, and RCTs—are formally about hypothesis testing. However, they are used to inform an ongoing interpretation of the change process, both within the part of the system that is specifically being explored (with an "intervention") and in terms of overall interactions within the system, and observation of broader processes of system dynamics.

III. Iterative Approaches

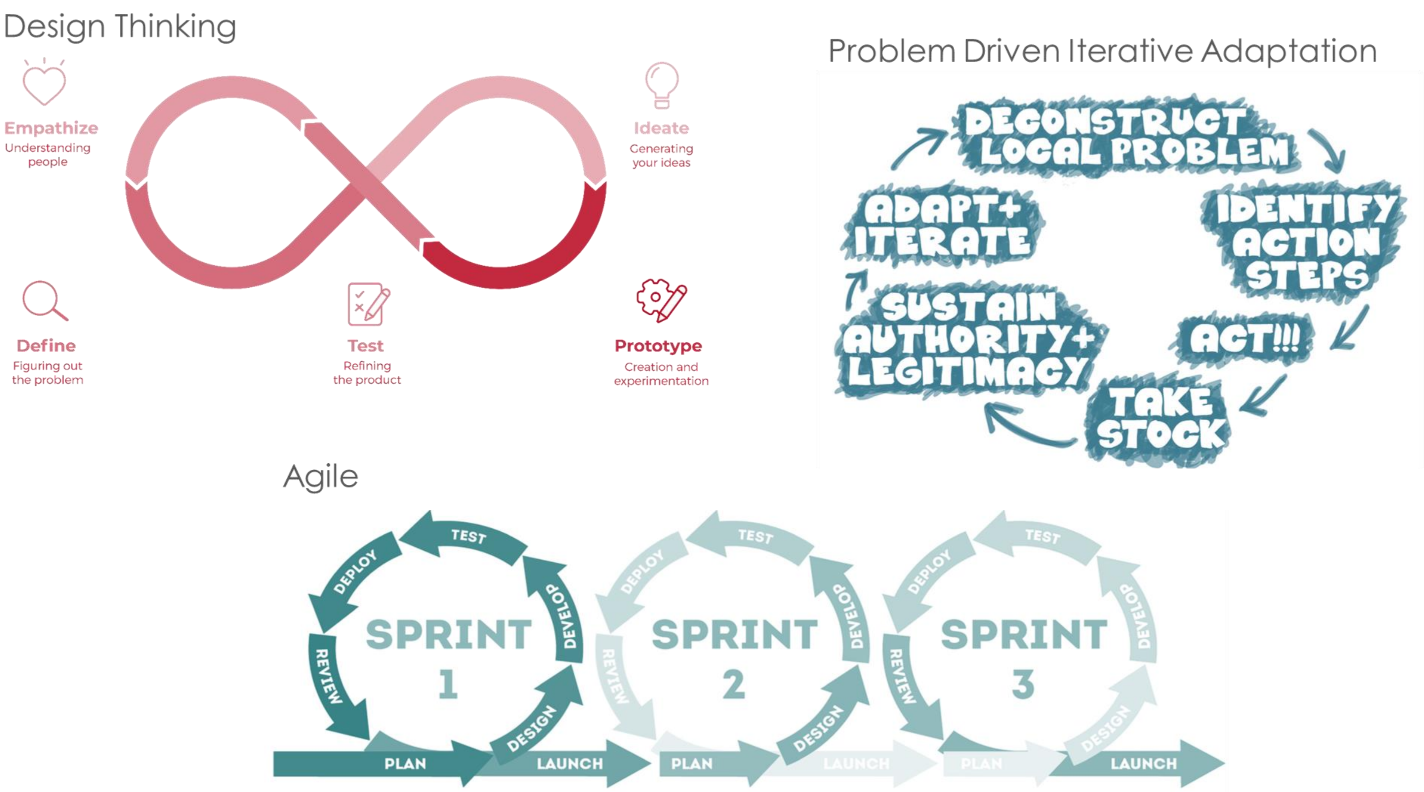
Iterative Approaches are a direct extension of the evaluation approaches just described and are particularly valuable when there is scope for incorporating feedback from the evaluative process into the design (see Section 3.2.1 Part III for more details.) They are central to the conception of an Adaptive Evaluation. They are designed for complexity—emphasizing flexibility and continuous learning as specific interventions and broader parts of a system experience changing behaviours, shifting contexts, and unstable environments. In this specific sense, iterative approaches are intrinsically 'adaptive'—they comprise the processes, tools, and methods that enable an intervention team to constantly adjust and refine the design in response to feedback.

Iterative approaches build from the theory of change to identify types of feedback that are valuable for the intervention team. In general, feedback needs to be collected on (1) the efficacy of the processes of the intervention, (2) the validity of the assumptions and theories that underly the intervention, and (3) the benefits to the end beneficiary. Processes can be visualized in flowcharts/**process maps** and tracked through project management tools and/or dashboards. The validity of assumptions can be tracked through both quantitative and qualitative techniques (such as quantitative surveys, administrative data analysis, and case studies). Benefits to the end-beneficiary can be measured by **user journeys**, **beneficiary interviews** and purposive surveys to measure impact, along with more formal **RCTs** and **quasi-experimental** approaches (such as a difference-in-difference and propensity matching)

Iterative approaches involve two complementary elements. First, there must be an organizational structure, relationship and mindset in the organization or team implementing the interventions that are open to documenting change, providing and receiving feedback and modifying the design. Second, there are specific techniques that can support this. The first is not a trivial requirement, since many implementing organizations are resistant to feedback (especially negative feedback) and display substantial inertia in designs. This is where the system diagnostic can help, exploring sources of leverage and blockages, and how to navigate through a complex stakeholder environment, in ways that allow feedback to work, at least in parts of the system.

Where there is scope for design responsiveness to feedback, specific techniques can be applied. Here there is an array of methods, many of them taken from adaptive management approaches originating in the private sector (**Figure 23**). These adaptive management techniques are aimed at learning fast and failing faster, to allow for rapid refinements that enable one to navigate through a complex system. Throughout, there is an emphasis on the importance of collaboration among multidisciplinary teams to allow for faster learning. There are a set of techniques that adapt **agile** approaches to design and implementation to the specific context of a development organization. This may involve **design sprints,** which typically involve a whole range of design thinking tools to specific tasks to produce testable prototypes, that can be taken to the field, with varying degrees of breadth (see next section on innovation and scaling). The approach of **problem-driven iterative adaptation** (PDIA) is essentially a version of this, adapted to groups of government officials who have the authorizing environment to explore innovations in the context in which they are working (Andrews et al, 2017).

Figure 23: Adaptive Management Approaches



Notice that each of these adaptive management techniques has a step for testing or reviewing or taking stock. An Adaptive Evaluation is designed to facilitate the testing and reviewing aspects of the adaptive management cycles. This involves (a) support for implementation teams in rapid design options for prototyping in short sprints, (b) frequent testing of what worked and measurement of desired outcomes (or more typically "outputs" in logic model language), for example with small scale A/B testing (that is typically a rapid randomization alternative treatments), and other relevant qualitative and quantitative techniques (including process tracing again, user journeys, beneficiary interviews), and (c) providing inputs to the implementation team on refining the design based on insights from this testing, to be assessed in subsequent rounds of prototyping. Beyond the initial design phase, there may be an intermediate phase of learning with more structure, for which observational analysis, positive deviance, randomization and more extensive process tracing can all be relevant.

An important aspect of an Adaptive Evaluation is that the evaluation and implementation teams work closely together. It is crucial for the success of any iterative approach that recursive feedback loops over short cycles are undertaken in close collaboration with the implementation partner to find effective intervention.

Consider a skilling and employment initiative to reduce unemployment, that involves, (1) a basic test to identify the skills and inclinations of a candidate, and (2) the ability to sign up for skills training. Feedback, in this case, would be about the number of people who took the test, got access to training, completed the training, and secured a job. Each of these is not trivial to measure. Suppose we have a mobile application to administer tests and make referrals to skills training institutes. In this case, tests and referrals to training could be monitored on a dashboard that uses the data from the app. Without a mobile application, tracking needs to take place via household surveys. Understanding whether people take up training could be done via random self-reported surveys on the app, or via training centres that partner with the skilling and employment initiatives. Finally understanding whether people who completed the training were able to get a job placement would also require either a random follow-up survey on the app or a merging of data from training centres that partnered with the skilling and employment initiatives. Even in this one example, we can see how this is a difficult exercise. Nonetheless, working to identify and collect feedback is the first step in being able to monitor that the intervention is working, and ensuring that there is a regular flow of information that allows one to learn about what is working.

IV. A Decision Map for Methodologies based on question type.

We began Section 3.2 by defining an Adaptive Evaluation—an approach and an associated set of techniques to answer complex questions. We then described the three main types of complex questions—system diagnostic, theory of change, and design responsiveness questions, and the three approaches to answer these questions—system-based, theory-based, and iterative approaches, respectively.

Often, in the process of an evaluation, it is difficult to know how to think about which of the gamut of methodologies to use, and in what situation. Standard impact evaluations typically stick to one methodology. Mixed-method evaluations, theory-based evaluations (e.g., realist evaluations) and developmental evaluations, emphasize methodological flexibility, but rarely specify how to decide which ones to use. By outlining approaches to specific kinds of complex questions, we have, in essence, articulated a way to think about which methodologies to use in an Adaptive Evaluation. The methodology used depends on the kind of complex question (system diagnostic, theory of change, and design responsiveness) and the type of question (interpretive or impact and effectiveness). **Figure 24** shows the decision-making map for methodologies based on the type of question. In Section 4, we will introduce two other ways to think about which methods to use—the nature of the problem (scaling or innovation) and the level of complexity.

Figure 24: A Decision Map for Methodologies

Types of Complex Questions

	1. **System Diagnostic**	2. **Theory of Change**	3. **Design Responsiveness**

Interpretive Questions	• Participatory System Actor Workshops – Cover Page
	  • Stakeholder Mapping
	  • Actor Mapping – Figure 15
	  • Identifying enabler and resistors of change
	  • identifying boundaries of the system
	• Circular Interviews – Figure 16
	• Roleplaying
	• System Dynamics Map (with feedback loops) – Figure 17
	• Ethnography
	• Case Studies
		• Logic Model – Figure 21
		• Fishbone/Ishikawa Diagram – Figure 18
			• Adaptive management process (agile, design-thinking, PDIA) – Figure 23
			• A/B Testing
			• Process Maps
			• User Journeys
			• Beneficiary Interviews
			• Needs assessment

Impact and Effectiveness Questions	• Flow of Intervention Map – Figure 19 • Circular Interview (for the intervention)	• RCT • Quasi-Experimental methods • Process Tracing – Figure 22 • Outcome Harvesting – Figure 20	• Process monitoring • Process Tracing • Circular Interviews
	Systems-based Approaches	**Theory-based** Approaches	**Iterative** Approaches

4. The Adaptive Evaluation Process: From Innovation to Scaling

In Section 3, we introduced an Adaptive Evaluation and its methodological elements. In this section, we propose an ideal sequence to answer complex questions. Throughout this section, we focus on innovation and scaling, two concepts that are central to development but are especially of interest to evaluators. Evaluations in development typically assess some change in policy, public action, or the system, which in general, can be usefully divided into two categories – *innovation* and *scaling*. 'Innovation' is the process of designing, refining, and adapting a creative idea in a local setting. 'Scaling' is the expansion of an innovation to the broader population of interest. The journey in development, from creative ideas to innovations, and from innovations to scale is inherently challenging and often involves changes in system behaviour or even system-level transformation. This process is neither linear nor predictable. It is no surprise then, that innovation and scaling are both often conceptualized as complex adaptive systems (see Section 2.2.2 for more details). An Adaptive Evaluation is designed to help support the transformation needed to foster sustainable change while adapting the techniques and methods to the varying levels of complexity through this journey.

Section 4 is organized as follows: Section 4.1 outlines a framework that maps techniques and methods based on the nature of the task (innovation and scaling) and the level of complexity; Section 4.2 describes the Adaptive Evaluation process, and Section 4.3 describes the evaluation journey.

4.1. A framework for Innovation and Scaling

Innovation and scaling are change processes that are integral to development. In Section 2.2.2, we introduced how innovation and scaling exhibit features of complex adaptive systems. Creative bursts spark ideas, some of which activate self-reinforcing positive feedback loops, resulting in manifold small-scale experiments that lead to successful implementation in a specific context (an innovation). Scaling-up is the process in which innovations are extended to reach a substantial part of the relevant population. This may involve extending an innovation in one part of an organization to other parts, the growth of the organization itself, or the replication of the innovation in other organizations (Cooley & Guerrero, 2016). It often involves taking an innovation to different geographical regions, that may have distinct characteristics and modes of functioning that were shaped and informed by their context.[19] And it typically involves more standardization, and adaptation to the initial or other organizations' broader culture, incentives, and operating procedures. As we emphasized above, scaling is itself a further process that needs to be underpinned by exploration and associated monitoring and Adaptive Evaluation processes, including what we described as design responsiveness

In Section 2.2.2 we gave an example of the trajectory of women's self-help groups in India that transformed from a powerful idea to a tried and tested innovation—largely due to pioneering efforts by MYRADA, a scaled innovation with successful implementation in several states in India—the early adopters being Andhra Pradesh, Kerala, and Tamil Nadu, and finally a central government programme—the National Rural Livelihood Mission. In Section 2.1.3 Part II, we explored how self-help groups were able to mobilize their Covid-19 response effort, culminating in 20,000 self-help groups in India, across 27 states producing 19 million masks, more than 100,000 litres of sanitisers, and nearly 50,000 litres of hand wash. This is yet another example of an idea—to involve self-help groups in the Covid-19 response—reaching a remarkable scale.

A large part of development is creating the conditions and environment to facilitate these emergent and self-organizing processes that transform ideas and take them to scale. Evaluations play a central support function, in aiding the measurement, testing and experimentation that takes an idea to scale, both to feed design and inform political and stakeholder support.

Standard impact evaluations in development frequently underestimate the complexity of the scaling-up process (even if their practitioners don't), effectively treating it as if it were a complicated problem (see Section 3.1.1 Part I for the simple, complex, and complicated framework). J-PAL, for example, believes ideas are taken to scale by a series of RCTs that determine what works (Glennerster, 2017). They often cite the example of Pratham, which conducted six RCTs en route its journey to scale a major intervention to tackle learning problems—Teaching at the Right Level. As we discuss further below, the RCTs were indeed important, at a series of key moments, but were only one part of the story of exploration and assessment within Pratham. The RCTs primarily helped confirm the effectiveness of their iterative adaptive approach and experimentation, rather than being the central driver of the innovation and scaling-up process. This becomes more apparent in the case of TARL Africa, an organization that aims to replicate Pratham's model to improve education outcomes

[19] This idea of innovation and scaling having emergent and self-organizing properties transcends far beyond development. For example, democracy was an idea, initially founded by the Greeks, that was experimented with in many different forms, and implemented successfully in some places (in the form of innovations). Eventually, over centuries, democracy scaled. As of 2019, more than half the countries are democratic (DeSilver, 2019). Nonetheless, each country has its own unique system and application of democracy. The same can be said about religions that scaled over centuries. They started as a thought, story or doctrine expressed in a text, and then spread, but different regions adopted it differently with their own unique interpretations and characteristics.

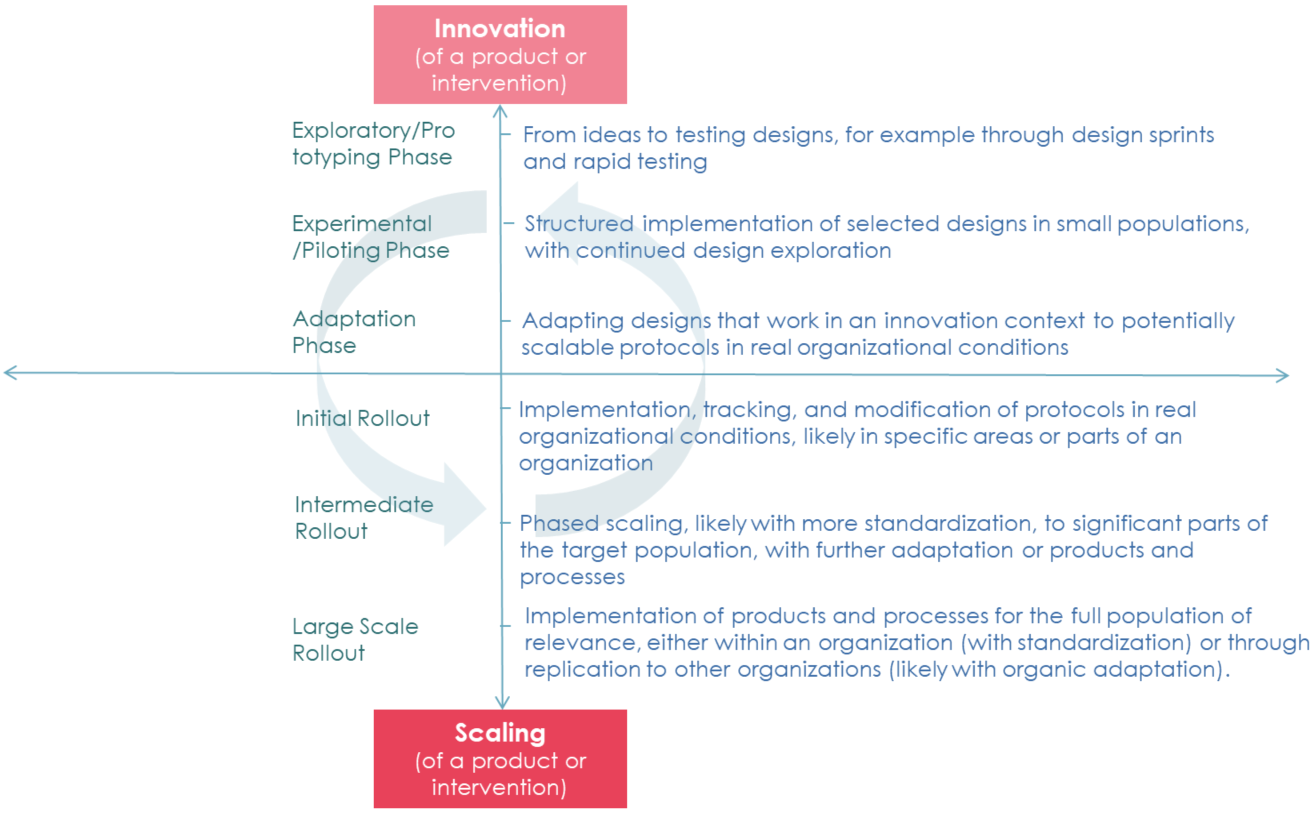
in Africa, as a collaborative venture between Pratham and JPAL. TARL Africa is also starting with a range of explorations with governments, learning mechanisms, and rapid small-scale experiments that were done by Pratham throughout. RCTs are best used to test ideas and innovations *after* these have been developed through an iterative learning and evaluation process.

An Adaptive Evaluation supports the facilitated emergence that is required to take ideas to scale because it inherently understands that scaling is a complex adaptive process. As is evident from the descriptions above, a huge component of this is experimentation, learning, and adaptation to ongoing changes, new contexts (including in the core organization) and geographical settings. An Adaptive Evaluation provides a set of tools and techniques that enable these learning mechanisms. The tools are flexible and depend on the level of complexity in the environment and the stage of the innovation. In other words, an Adaptive Evaluation understands that RCTs are valuable in certain parts of the scaling-up process but is aware that no one methodology can aid a complex process like scaling.

Section 4.1 first introduces the innovation-scaling continuum, and then the complexity continuum. Finally, it plots methods onto a 2-dimensional space featuring both continuums, clearly outlining what methods to use at different stages of the scaling-up process.

4.1.1. The Innovation-Scaling Continuum

Figure 25: The innovation-scaling continuum



Innovation and scaling can be viewed as a continuum. We begin with the highly exploratory innovation stage in which there are several potential designs and possibilities of interventions. These possibilities are informed by descriptive observations and existing interventions. For example, consider an intervention to increase test scores in rural India. The process of innovation begins at a tiny scale, say in a neighbourhood of ten families that go to one tiny school in a sparse, remote village. We then quickly enter a rapid prototyping stage which initiates the iterative learning process. Slowly, we begin standardizing/concretizing some learnings based on what worked, while simultaneously innovating on whatever remains to be refined, narrowing down on the design and possibilities. As we begin to standardize more features of the intervention, we also expand our scale, say now to fifty families and three large village schools. At some point, we reach a stage where a design is robust in this context, and ready for early-stage rollouts, across many schools or selected schools in many school districts. This might use the language of "protocols" for testing, or "minimum standard intervention/product." As we encounter and learn from, and address implementation problems in such an initial rollout, we move to a broader expansion or intermediate rollout, then a full-scale state rollout, and finally, a national rollout, or even an international rollout.

It is worth pointing out a few key observations as one traverses from innovation to scale in this continuum (from top to bottom in **Figure 25**). First, there is experimentation *throughout* the continuum from innovation to scale—what changes is simply the nature of experimentation. In the various levels of innovation (exploratory, experimental/piloting, and adaption), there is usually more experimentation on the design features of the intervention. In the various levels of scaling (initial rollout, intermediate rollout, and large-scale rollout), there is typically more experimentation on process and implementation.

Second, in the innovation space (the top half of **Figure 25**), the experimentation takes place at a smaller scale with fewer stakeholders involved. In general, this implies that (1) data collection and other observational information are more likely to accurately reflect what is happening, even if they are not statistically representative of the population (2) implementation processes are easier to control and manage, (3) tweaks in response to feedback can be made relatively quickly at a low cost, and (4) the time to detect effects of a change may be relatively small, especially in comparison to the scaling space (the bottom half of **Figure 25**). As a result, experimentation at early levels of innovation is relatively easier to design and run (albeit with the caveat that many effects can only occur and be observed over longer periods). In the scaling space (the bottom half of **Figure 25**), the experimentation takes place at a larger scale, with multiple departments and stakeholders involved. This comes with its own set of systemic challenges. In general, in comparison to the innovation space, the large scale implies that (1) measurements are less accurate (they are noisy and, may be prone to manipulation owing to incentives of actors), (2) implementation processes are difficult to monitor and control for quality—there may be issues of lack of ***fidelity*** in implementation, (3) tweaks in response to feedback can be made but slowly, and at a much higher cost, and (4) the time to detect effects of a change is typically larger, given it needs to trickle-down or trickle up to various levels of a system. **Figure 26** below illustrates the difference in challenges between innovation and scaling. RCTs are suitable for some later levels of innovation but are typically less suitable for scaling, which comes with much more complexity.

Figure 26: Implications of Innovation and Scaling

	Innovation (Exploratory Piloting Phase)	**Scaling** (Intermediate/Statewide Rollout)
Scale	Smaller • Operate in smaller geographical units • Involve smaller teams and fewer stakeholders	Larger • Operate in much larger geographical units • Involve departments and many stakeholders • Involves working with the system and may involve system change
Ability to Tweak	Faster, Lower effort, Lower cost	Slower, Higher effort, Higher Cost
Implementation	Easier to monitor and control for quality	Much more difficult to monitor and control for quality
Measurement	More precise	More noisy and prone to error (or even manipulation). Typically involves administrative data from government
Time to detect effects	Effects of changes in design are realized reasonably quickly	Effects of changes in design are realized over longer periods of time
Learning	Iteration cycles are shorter, so learning is rapid	Iteration cycles are longer so learning is slower
Evaluation Features	**Mixed**, relying more on **quantitative** **Theory-based** hypothesis testing **Rapid prototyping, experimentation, measurement, iteration** to improve intervention design	**Mixed**, relying more on **qualitative** **Theory-based** hypothesis testing **Periodic experimentation, measurement and iteration** to understand gaps and system failures **Participatory** to account for multiple stakeholders **System Based** to account for multiple levels of interaction **Process- Oriented** to account for implementation hurdles

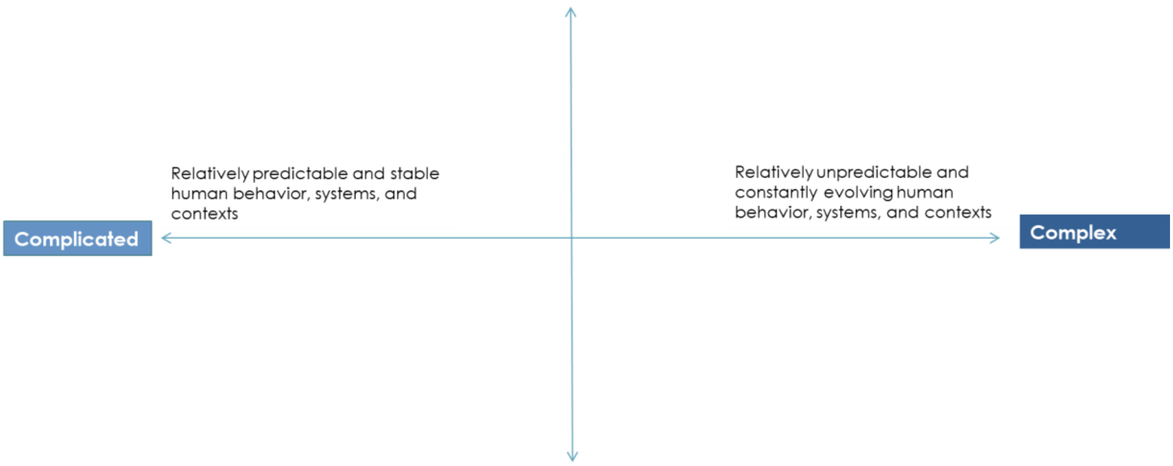
Third, and perhaps most pertinently, we recognize that innovation and scaling are not necessarily linear through the steps. For instance, an evaluation may start at an intermediate rollout only to realize it is not working—then scale back and restart the innovation process again. Moreover, while the innovation-scaling continuum might suggest that one can move seamlessly across the levels, in reality, this is messier and much more interesting. It is best to think of each level in innovation and scaling as a circular, iterative, nonlinear learning feedback loop, similar to the ones described in **Figure 23** on adaptive learning processes.

4.1.2. The Complexity Continuum

Complexity can also be viewed as a continuum (see **Figure 27** below). We define the extent of complexity in relation to the multiple features of complex systems discussed in Section 2.1, that is the degree of openness, nestedness, non-linearity, dynamism, unpredictability, cyclicality, self-organization, emergence, and adaptation in a complex adaptive system. As argued in Section 2.2 development is intrinsically a form of a complex adaptive system. However, there are varying degrees of complexity with respect to these features, that in turn influence the nature of the problem, and how far an evaluation needs to follow an adaptive approach to be effective. A high level of complexity corresponds to the "complex" region in **Figure 27**, and a moderate level of complexity corresponds to the "complicated "region in **Figure 27** (note 'complicated' comes from Cynefin Framework explained in section 3.1.1). Note also that complexity can be a feature of the intervention structure as well as the context in which it is working.

Evaluations conducted in the complex region of **Figure 27** (think of a hurricane response or the initial phases of COVID-19's spread) have many more challenges relative to the complicated region. Data is more limited, quality is more suspect [20] (i.e., real-time, granular data is not as abundant), and it is difficult to randomize into treatment and control (due to multiple and changing interventions, contamination, and shifting conditions). Moreover, cause and effect relationships are non-linear and muti-directional. Evaluations conducted in the complicated region of **Figure 27** have slightly more predictable and stable environments (for example, a computer interface in which a new app is released). They are, in general, slightly less challenging, data quality is better and of greater volume (real-time, granular data is more readily available), and it is relatively easier to randomize into treatment and control. **Figure 27** below illustrates the complexity continuum.

Figure 27: The complexity continuum



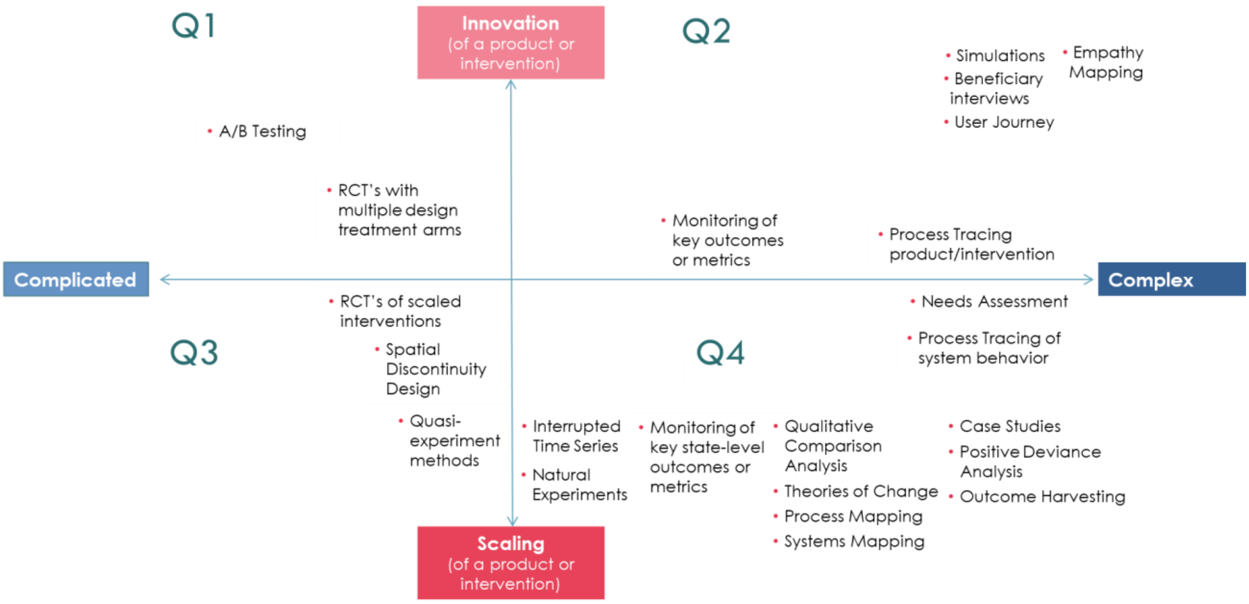
4.1.3. A Decision-Map in the Innovation-Scaling and Complexity Continuum

In **Figure 28** we develop a heuristic map of techniques on the two dimensions elucidated above— (1) innovation-scaling and (2) complexity (see **Figure 28** below). The main idea is that in an Adaptive Evaluation, we use different tools depending on the complexity of the system and the nature of the task in the scaling/innovation continuum. This complements the decision map of techniques introduced above, which was based on the **type** of questions to be answered (see **Figure 24** above).

This is only illustrative, and in reality, different techniques occupy a range in the complexity and innovation-scaling space. For example, there are RCTs in less complex environments and RCTs in more complex environments, but overall, they can handle less complexity than process tracing, which tends to be more flexible even when data is scarce or there is no counterfactual to exploit. Finally, the list of techniques mapped is not meant to be exhaustive, but just an illustration of typical methods that are suitable for each quadrant.

[20] This is of course a generalization, and there are a few examples of high-quality abundant data in very complex settings. High-definition satellite data can provide real time, granular information about land quality, and can be used to predict poverty. In addition, governments are becoming increasingly better at collecting and analyzing big data. Nonetheless, despite advancements in technology, in general, high complex environments have fewer such sources of data that are easily accessible and useable.

Figure 28: Mapping techniques to the complexity and innovation-scaling space



A. Complicated and Innovation

We begin with the first quadrant. Here the relatively lower level of complexity may allow for some comparator cases that can be randomized. At the early stages of innovation that fall in this area, when there are many untested design options, and we are in a relatively lower complexity environment (e.g., in app development), we might use A/B testing (that is, a simple form of RCT comparing randomized alternative treatments). As we have much more refinement in design, and are at a more advanced stage of innovation, in which we have narrowed it down to, say, 2-3 choices, we can use more developed RCTs, typically with much more data collection on participant characteristics and process monitoring, implemented at a larger scale than the first pilots. More developed RCTs, for example with survey data on household or firm characteristics, can handle more complexity than A/B testing and are often conducted in social field experiments.

B. Complex and Innovation

Now consider the second quadrant. Here the high level of complexity rarely allows for comparator cases, let alone randomization. At the early stages of innovation, when there are many possibilities, and when complexity is very high, we might use beneficiary interviews, empathy maps, or other qualitative techniques, to narrow design choices. Once we have more of a running minimum viable product, we can consider user journeys. Economic modelling may be useful, based on existing conceptual and empirical work, to provide structure, including running of simulations if these have quantitative parameters, though these should remain in conversation with the qualitative work. At a more advanced stage of innovation, we can consider systematic process tracing and logical tests of alternative hypotheses, provided there is sufficient quality of data for (at least) some proxies to the outcome of interest.

C. Complicated and Scaling

The third quadrant involves scaling in a relatively lower complexity (consider distributing bed nets to school children all over the country in a period of economic, social, and political stability). In a relatively lower level of complexity, it is likely that there are good comparator cases without the scaled intervention (that is, a good counterfactual). As a result, for early-stage rollouts of an innovation at the protocol phase (say in the district), RCTs, or quasi-experimental methods, can be useful instruments, where these are feasible. For a state-level rollout (consider a state policy for minimum wage in stable economic conditions), which is a more advanced stage in scaling, there may be scope for spatial discontinuity design or a quasi-experiment (e.g., difference-in-difference, RDD, and instrumental variables).

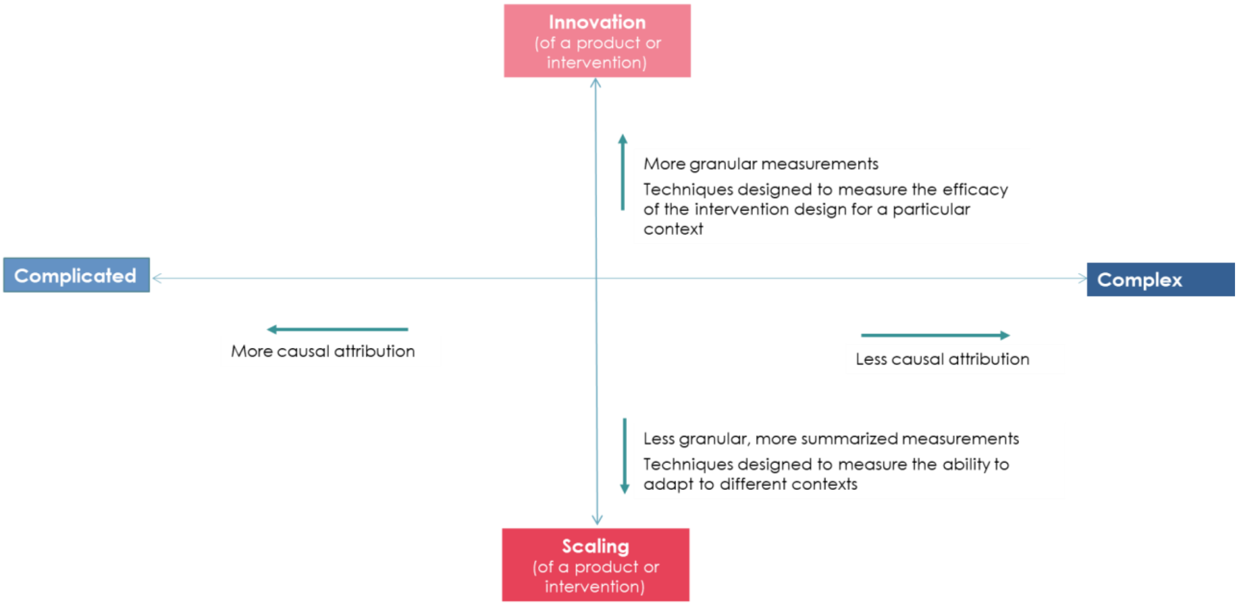
D. Complex and Scaling

Finally, let us examine the fourth quadrant (a good example is managing protocols that seek to affect population behaviour during COVID-19), which is in many ways the most challenging to think about, but is also the space many complex questions inhabit. Let us start at a very high level of complexity and a large-scale state or national rollout. In this space, we can use a variety of eclectic mixed method techniques, from case studies and positive deviance analysis (to interpret variation) to process and systems mapping, and outcome harvesting for interpretation of multiple processes. Monitoring of state-level indicators is important for interpretation, provided_quality data can be collected. This can support the application of interrupted time series and natural experiments. Especially in the early-stage rollouts in very complex environments, one would typically conduct a needs assessment to understand the resources needed to scale, and process tracing of system behaviour.

E. Implications of the framework

There are three important aspects worth noting in the framework. First, there are clear trade-offs present along the two axes (see **Figure 29** below). As one moves from lower to higher complexity, the causal attribution of the techniques becomes increasingly less precise. In **Figure 29** for example, in the *complicated* half of the diagram, most of the methods are RCTs or quasi-experiments, which are most powerful for the statistical assessment of causal attribution. On the other hand, in the *complex* half of the diagram, most methods are qualitative or involve qualitative interpretations of observational quantitative data, which have reduced causal attribution but are more powerful at interpreting complexity. Similarly, as one moves from innovation to scale, the measurements are less granular and more summarized, and techniques are designed to measure the ability to adapt to contexts. As we have mentioned before, scaling typically involves replicating an innovation in different contexts, whether within a state or beyond, and this requires a systematic assessment of how context influences the effect of interventions.

Figure 29: Tradeoffs as one moves across the innovation-scaling and complexity space



Evaluators need to be cognizant of these trade-offs when deciding the tools one should use in an evaluation. An Adaptive Evaluation is acutely aware of these tradeoffs. It attempts to make the choice that best fits the nature of the task (innovation and scaling) and the complexity of the environment, cognizant of how the environment changes over time. An evaluator has to **adapt** appropriately to the new levels of complexity.

Second, notice that tools in higher levels of complexity can also be used in lower levels of complexity (that is, all tools in Q4 in **Figure 28** can be reasonably used in Q3, with the caveat that they have less causal attribution). However, the reverse is not possible—tools used in lower levels of complexity (the "complicated" region") are inadequate at higher levels of complexity (the "complex" region). An RCT may not be feasible at higher levels of complexity, for example, because randomization may not be possible, or worse, there may not be a counterfactual or comparator in the first place.

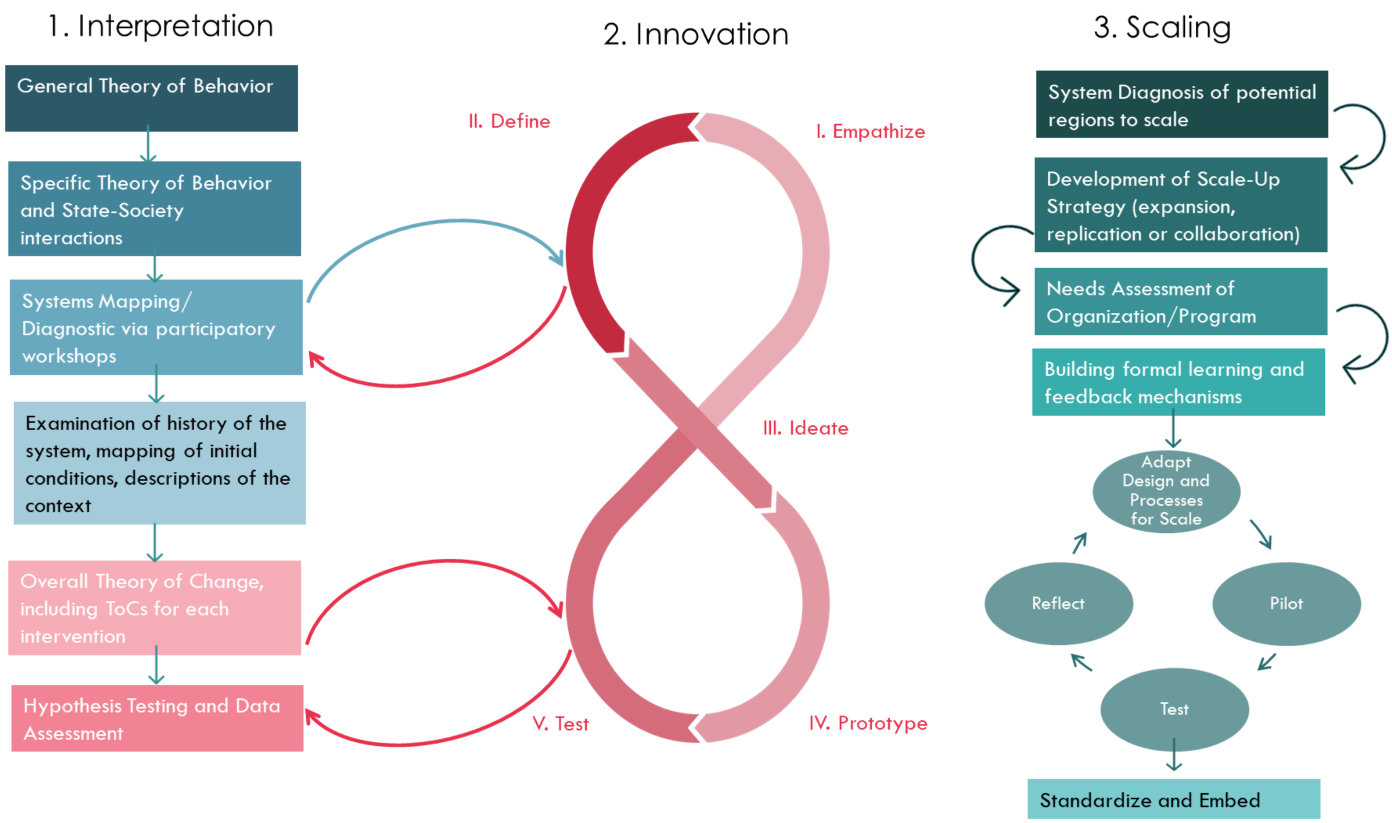
Third, **Figure 28** offers another decision-map that can help inform which tools and techniques to use, based on the nature of the task and level of complexity. The fact that an Adaptive Evaluation is itself adaptive in its prescription of tools and methods, is a key distinguishing factor relative to other evaluations.

4.2. The Adaptive Evaluation Process

So far, we have further conceptualized innovation and scaling and provided another decision-map to understand which technique or methodology one should use, based on the nature of the task and the level of complexity. What then would an evaluation process look like over the various phases of change? In this section, we explore a typical Adaptive Evaluation process.

An Adaptive Evaluation has three main phases—interpretation of system and processes, assessing innovation, and scaling. Together, these phases provide (1) an evaluation framework to understand the role of systems, context, history, and theory, (2) an evaluation framework to assess existing change processes and support rapid measurement, prototyping and experimentation, and (3) a further evaluation framework to support scaling. **Figure 30** below illustrates the three phases of an Adaptive Evaluation.

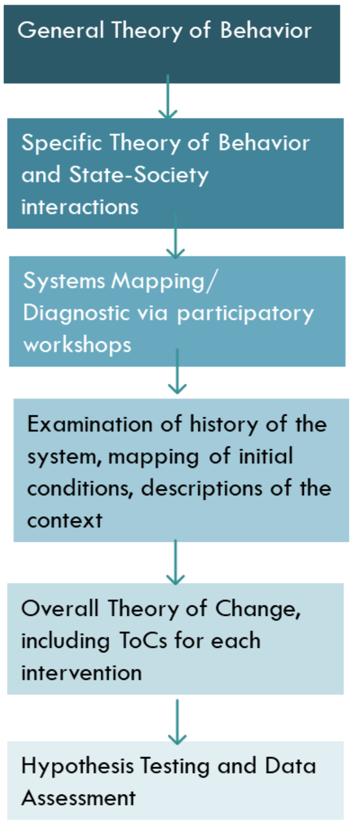
Figure 30: The three phases of an Adaptive Evaluation



Let's consider an "ideal" sequence for a complex problem, for which "we" (the implementing agency, leaders etc) want innovations that can go to scale. We suggest a life cycle. It begins with an understanding and interpretation of the relevant theory, history, context, and system, including the theorization of change processes in ways that can be taken to data assessment. Then it moves to a phase of systematic learning, in which interventions, purposively selected keeping in mind the complex landscape, undergo cycles of improvements and refinements in design and process, based on regular feedback. Finally, once an intervention, or set of interventions, has been tried and tested within a particular setting, through a process of iterative learning, it is launched for scale. This involves an assessment of the new landscape and a different learning cycle. Of course, in development, the 'ideal 'sequence, may not be possible, and actual evaluations will likely change the order and sequence based on the specific demands of the task and changing needs of the environment.

4.1.1. Interpreting the system and change processes

Figure 31: Phase 1- Interpretation



Interpreting the system and change processes is the first phase of an Adaptive Evaluation. This phase is dedicated to answering a range of interpretive questions—descriptive questions that provide an account of the history, evolution, theories, and characteristics of the complex adaptive system the development problem resides in. In particular, this phase addresses two of the three kinds of questions elucidated in Section 3.2.1—system diagnostic and theory of change questions. Interpretation includes both system-based and theory-based methods for change. This phase forms the basis of understanding for subsequent stages of an Adaptive Evaluation.

The interpretation phase begins with an understanding of the general or overarching theories of behaviour around the problem or desired change. Consider a social enterprise whose desired change is to improve the learning deficits of children at primary school. General theories could be theories about early childhood development, role models, modes of learning, and pedagogies. From this general theory, we can move to specific theories. In the example of learning deficits, this includes theories behind current policies or interventions, theories about the supply of schooling (are there enough classrooms, teachers, and learning materials) and theories about the demand for schooling (are the returns to education high and are the costs to education low enough). Understanding the general and specific theories of behaviour provides a lens with which to investigate the problem.

This initial theorization is accompanied by or embedded within, a full system diagnostic, in recognition of how development is inherently a complex adaptive system. System-based approaches, such as actor mapping, roleplaying, circular interviews, and system dynamics, among others, can be used at this stage. These tools and methods, described in Section 3.2.2 Part I, will help (1) paint a picture of the non-linear relationships that exist in the system, (2) identify positive and negative feedback loops, (3) understand enablers and resistors to the desired change, and (4) point out factors outside the system that affect the desired change. A system diagnostic that addresses learning deficits at primary school would comprise a full analysis of the nexus and interconnections between teachers, students, parents, the schools, the education ministry, and so on.

Finally, in acknowledgement of the dynamic nature of systems (see Section 2.1.2 Part IV) in development, we complement the system diagnostic, with an examination of the history and initial conditions. Complex adaptive systems are known to have a high sensitivity to initial conditions. Understanding initial conditions ensure evaluators and implementers are aware of the environment an intervention resides in. Perhaps more pertinently, it helps grasp the dependence of an intervention on its initial conditions, and how a change in these conditions in the future might affect the intervention. Initial conditions can be mapped using publicly available datasets over time (including administrative and geospatial data), that present a full picture of the state of deprivation in the region of interest. For our specific example of learning deficits at primary school, this would involve mapping the trajectory of learning outcomes over time and understanding different policy shifts in education to date.

After these steps, one can begin identifying potential interventions, keeping in mind, the high-level theorizations, the specific theories on state-society interactions, and the blockages in the system. A potential solution candidate to address learning deficits in primary schools, for example, should be (1) based on some theories about learning and childhood development, (2) designed to address either demand or supply side constraints in education, (3) bypass or neutralize blockages in the system, and (4) learn from past successes and failures.

From the system diagnostic and interventions, we continue working on the theory of change, this time, a general theory of change to affect the desired change within the system, and then a family of theories of change for each potential intervention. Theory-based approaches, described in Section 3.2.2 Part II, are central to this part of the sequence. In the case of learning deficits, this would map how and why, a particular intervention candidate, for example, mobile tablets and specific pedagogy, will lead to an increase in sustained attention and thus learning outcomes for primary school children.

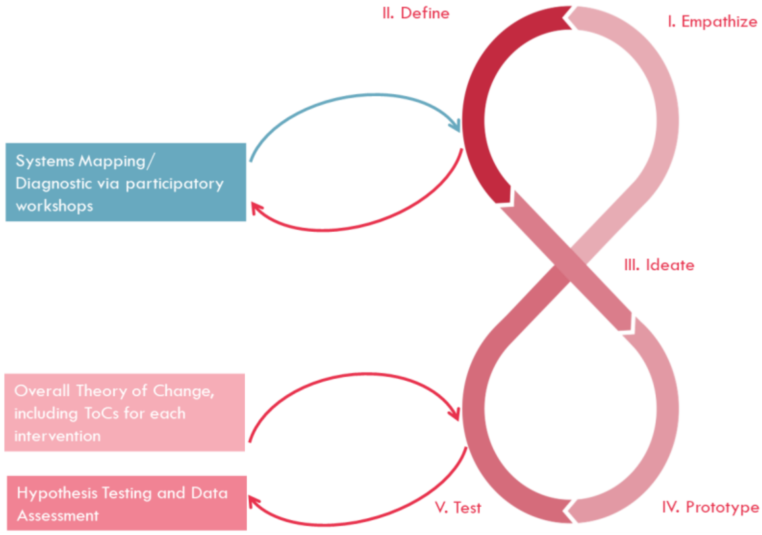
Finally, at this stage, we test the theories and assumptions underpinning each of the intervention candidates, to get a better understanding of which of the battery of interventions seems most feasible. This will be complemented with a review of the available sources of data (i.e., the data assessment). This could include administrative data collected by the government, available surveys conducted by civil society organizations and/or private firms, and tailored surveys for the interventions. The data sources will be identified in relation to specific hypotheses derived from the theory of change exercises, but also keeping in mind information that might be valuable for innovation or scaling.

Two features are worth noting in this phase. First, the sequence intentionally moves from general to specific, slowly building to a detailed and concrete understanding of reality that is specific to the intervention. This is by design. It ensures that even as we finetune and tweak specific aspects of an intervention (see the next section on the innovation phase), we have an understanding of the complex forces in the bigger picture and are cognizant of these interactions throughout. Second, we oscillate between theory and system-based approaches in the sequence. We see this as important because we want theory to be informed by an understanding of the systems, and our systems work, especially on the flow of interventions, to be guided by theory.

4.1.2. Innovation

The innovation phase follows the interpretation phase and addresses design responsiveness questions. This is where the most promising interventions undergo cycles of prototyping, testing, and refinement to improve the design of the product (assuming, of course, that the implementation team has the culture and practice of working this way). **Figure 32** has a sequence adopted from design thinking, although other similar non-linear, iterative processes can also be used. *Empathize* involves really understanding the problem that one is trying to solve more intimately, through the lens of the end beneficiaries and other key actors, including front-line and higher-level workers. Roleplaying is an apt example of a technique to emphasize, although this can also be done through in-depth personal interviews. An important part of this step is listening to the needs and concerns of those closest to the problem. *Define* is about clearly articulating a problem statement that captures all findings from understanding the position of the relevant actors. *Ideate* is a creative phase in which one carefully looks at the solutions identified that can resolve the problem at hand. In this part of the innovation phase, one looks to think out of the box and challenge assumptions to come up with a more detailed solution. During *prototyping*, the ideas are brought to life and implemented. Finally, during *testing,* the prototypes are tested to see if they work. What isn't working is analyzed, and used to redefine the problem, and the cycle begins again. A host of iterative approaches, as outlined in Section 3.2.2 Part III, can be used here. Note that the process of prototyping is used to learn and update previously done theories of change and system diagnostics undertaken in the interpretation phase (see arrows in Figure 32).

Figure 32: Phase 2- Innovation

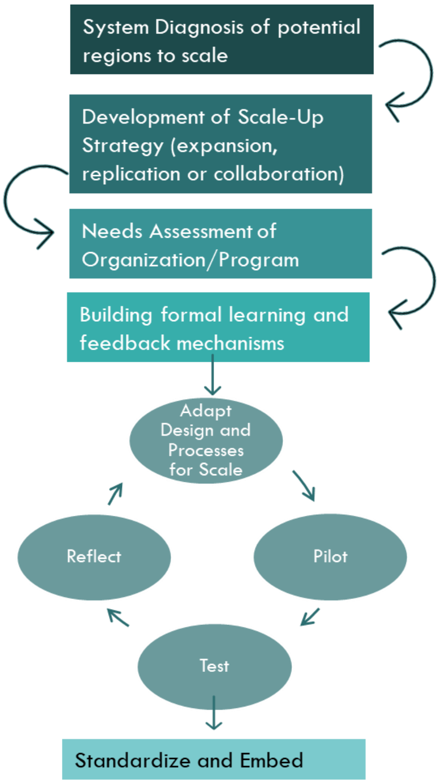


This phase of learning and adapting is essential to the success of any innovation, especially in development, where they are so many complex forces at play, that it is near impossible for a solution to succeed without learning about what works and doesn't. A learning mindset and culture require discipline and need to be nurtured. It most certainly cannot be assumed. Agile management techniques help facilitate and guide these processes with a structure that helps capture learnings, and their application in improving the intervention.

The outline here is of an ideal process with an implementation partner who has the culture, incentives, and practices to experiment and explore. Many organizations are not like this, perhaps especially, but by no means exclusively, in government agencies. In these cases the task is related but different: part of the system diagnosis concerns how feedback occurs, and designs change in **this** organizational context, taking the spirit of innovation and learning to the organizational and authorizing environment that actually exists.

4.1.3. Scaling

Figure 33: Phase 3 - Scaling



Scaling is a critical phase of the development process. Adaptive evaluation is a complement to this—and is particularly relevant because of the intrinsically system-wide feature of the scaling process. Scaling comes after an intervention has been prototyped to some level. The Adaptive Evaluation structure is essentially the same as for the innovation phase, but with different objects of analysis. The first step is a system diagnosis of the scaling question: is it, for example, an extension of an innovation throughout an organization (a ministry, public or private enterprise), an extension to the whole population within a region, or taking the innovation to new regions and countries. The characteristics of the end beneficiary and the functioning of the key stakeholders, entities and processes can significantly transform the problem in new regions. A system diagnostic helps chart these differences.

Based on an understanding of resistors and enablers to change in different regions or organizations, a scale-up strategy is formed. Scaling up can happen through ***expansion***, which refers to the growth of the current operations, ***replication***, which involves the adoption of the same concept by the government or the market, and ***partnerships***, which involve a combination of expansion and replication (Cooley & Guerrero, 2016). Deciding which is the best way to scale is an important step.

A needs assessment is then undertaken to understand the gaps that need to be filled within the organization or by the system to be able to scale. Finally, one needs to understand the existing formal learning and feedback systems, and what will be needed to enable an intervention to adapt to the demands of the new population and implementation context.

With this preparation, we enter another iterative phase of adapting design and processes for scale. This may include further piloting, testing, and reflecting to further adapt the design, or, in some cases, going straight to scale! Feedback is collected to ensure that processes and outcomes are sustained in the new contexts. Iterative approaches outlined in Section 3.2.2 Part III are useful here, though with very different organizational processes than outlined above, since these need to "work" within the incentives and structures of the new context. This phase ends with a probable further standardization of the process, embedding it into the implementing organization's protocols or standard operating procedures.

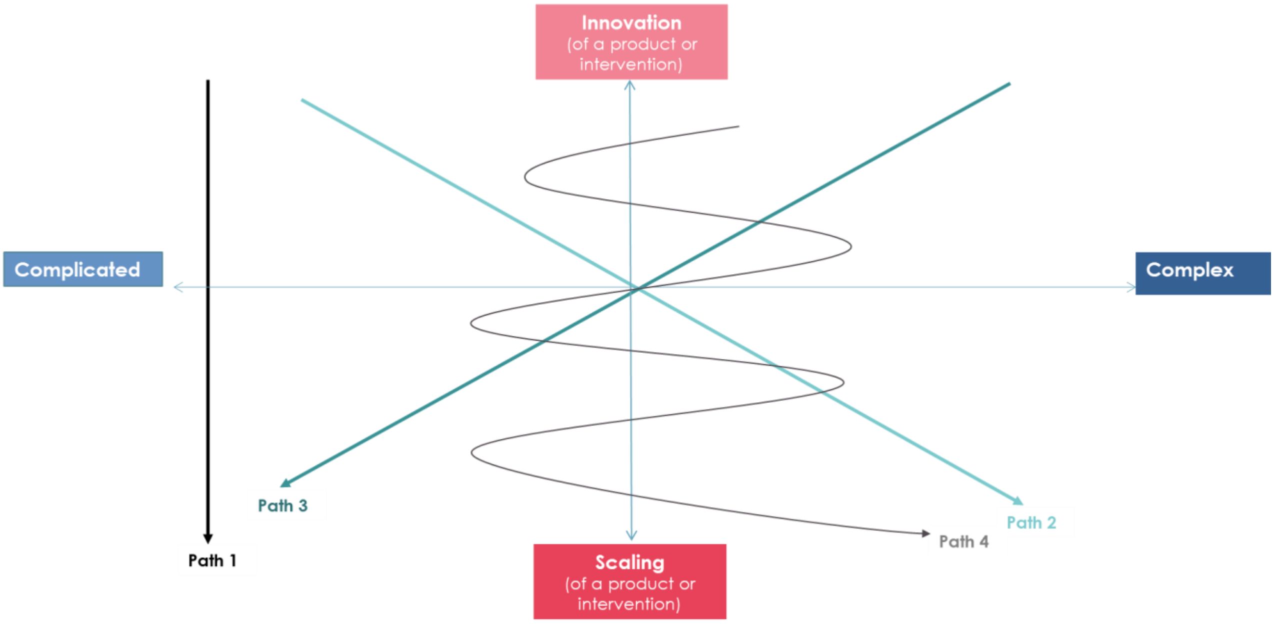
The scaling process is arguably the most difficult because it involves working through the levels of the system. Iterative learning does not stop with the innovation process, since new issues emerge during scaling, whether in terms of organizational behaviour—new parts of the system reacting, resisting, accelerating change or following unintended pathways—or a whole range of unpredictable human and social responses. So, while going to scale almost always involves crystallizing intervention approaches into protocols that can be supported by the financial and organizational resources of the system, the exploration, evaluation, and adjustment of designs continues.

4.3. Evaluation as a Journey

Until now, we have examined three different ways to determine relevant tools to use in an Adaptive Evaluation—(1) the kind of question (see **Figure 24**), (2) the level of complexity (see **Figure 28**), and (3) the stage in the innovation-scaling continuum (see **Figure 28**). These decision-making maps are useful because they understand that evaluations, much like the complex adaptive systems in which it operates, have their own unique journeys. Evaluations need to be able to respond with tools, techniques and methodologies that are appropriate to the path or lifecycle that an intervention finds itself in.

We can map evaluation journeys in the innovation-scaling and complexity space (see Section 4.1 for an introduction to this space) using illustrative time vectors. We begin with some purely theoretical trajectories over time, to illustrate (see **Figure 34** below). An ideal pathway for an intervention might be thought of as moving sequentially down from innovation to scaling for a given level of complexity over time (see Path 1 in **Figure 34**). This is rarely the case. The environment or the intervention designs may become more complex with time (say moving into a recession or entering a pandemic or starting an international conflict), and one can imagine moving diagonally and downwards and to the right across the diagram (see Path 2 in **Figure 34**). Or the problem becomes less complex, for example, when explorations and experience lead to more structured and predictable relationships across different parts of the system. This occurs when an extensive policy experiment process converts a complex problem into a complicated one. Here one can imagine moving diagonally and left with time (see Path 3 in **Figure 34**). We earlier suggested that this was a feature o Progresa (later Oportunidades and Prospera), the successful cash transfer program in Mexico (see Section 3.1.1 Part II B). Most evaluation journeys, much like intervention journeys are impossible to predict and will follow a more exploratory path. These are due to a variety of reasons, including changes in governments, leadership changes, shifts in context, emergencies or crises or, importantly, changes in questions, experience, and understanding as the exploration and evaluation unfolds. Path 4 suggests one "typical" journey, which starts in innovation and complexity, explores it in ways that reduce complexity, but then discovers new levels of complexity, as it starts scaling, and so on, as the process continues. Overall **Figure 34** illustrates how every evaluation starts at a particular level of complexity, scale, and time, and has its own journey over time. The evaluation tools must adapt to respect the requirements of time, scale, and complexity.

Figure 34: An evaluation can follow many paths over time



Let us consider a real-life evaluation journey—the journey of Pratham's search for an intervention that would tackle deficits in basic learning amongst Indian children — which was already introduced above. Pratham is an Indian NGO which was founded in 1995 to improve education in the slums of Mumbai. It has grown to be one of the largest NGOs in India, with a central goal of "every child in school and learning well". Pratham developed an intervention, now known as Teaching at the Right Level (TaRL), that is globally recognized, and is also being implemented in several African countries (under the TaRL Africa organization) and by a group working with state governments in Brazil. In its essence, it involves an *organizational* modification to practices in primary schools, that tests children in reading and math, and then devotes special sessions to an interactive *pedagogy* that effectively brings the children up to the appropriate grade level.

This, however, took many years of exploration and discovery that looks very like Path 4 in Figure 34! It turns out that helping children catch up with basic learning skills is a complex problem. Pratham's own learning started with tutoring in Mumbai and continued with the effective search and continual assessment of techniques for rapid, low-cost testing of children's reading and arithmetic skills (the ASER assessment tool) and applied pedagogies for catching up. It was discovered that the pedagogy was even effective when implemented by volunteers with only rapid training.

But to be effective at scale it was important to influence teaching in the state systems of India. This was a whole other layer of complexity, involving convincing education leaders of the problem, and engaging with bureaucratic school systems across India (primary education is a "state-level" subject in India's federal system). Pratham followed multiple paths. At the national level, it developed an extraordinary, statistically representative survey of basic learning schools, involving testing of children in their homes by an army of volunteers from many organizations, managed and analyzed by Pratham. This was key to developing the consensus that learning (as opposed to attendance) was and is a major problem. Then Pratham worked with individual state governments to design the organizational changes (supported financially by individual and philanthropic donations.) This also took some years of experimentation to crack through two designs. The first involved securing bureaucratic authorization, rules, training and specific direction for the front-line state workers in rural areas, and the array of printing of materials, training of teachers and instructions to principals. The second involves short-term "camps" with intensive use of the pedagogy by teachers trained in the technique. This was a long-term, and continuing, journey.

What is particularly interesting is that Pratham invited JPAL to accompany them on this journey, with some 6 RCTs, covered in four academic papers (Banerjee et al., 2007; Banerjee et al., 2010; Duflo, 2015; Banerjee et al., 2016).[21] While each of these provided rigorous, statistical assessments of the intervention, we argue that they need to be placed in a broader context. Pratham is profoundly a learning organization, shaped by its leadership, culture and practices. The RCTs need to be seen as just one part of a distinctly Adaptive Evaluation pathway, in which exploration and testing of designs are continuous, using rapid quantitative and qualitative techniques, which are periodically selected for traditional impact evaluation (not least for the reputational gains vis-á-vis an often-resistant authorizing environment).

5. A Concluding Comment

In this paper, we introduced Adaptive Evaluation—a complexity-based approach to systematic learning that is designed to support innovation and scaling in development. The underlying motivation behind an Adaptive Evaluation is twofold. First, standard impact evaluations in development, while useful and informative, are ill-designed for interpreting change within a system, even less for system-level transformation. This is due to a variety of reasons, including, among others, an emphasis on testing over learning, their focus on parts of a system rather than the whole, their stress on linear causal relationships, and their insistence on privileging a narrow set of tools (namely, RCTs and quasi-experimental methods). This makes standard impact evaluations less flexible, and unable to capture and absorb the complexity intrinsic to innovation and scaling. Second, while some other evaluation approaches are designed for complexity and encourage learning and methodological flexibility, they do not specify which techniques to use, and when and how to use these techniques. Moreover, existing complexity-based evaluations typically emphasize only one aspect of complexity— such as innovation (as in a developmental evaluation) or the role of context in the realization of an outcome (as in a realist evaluation)—but seldom combine more than one aspect. The Adaptive Evaluation approach developed here is an attempt to provide a clear decision-making roadmap on the type of techniques to use, while maintaining a more holistic approach to complexity—one that honours the role of history and context, facilitates innovation, and most importantly, also supports scaling.

In the previous sections, we outlined in detail the questions, methodological elements, and processes of an Adaptive Evaluation. We want to conclude with a final comment on for the broader role of adaptation. An Adaptive Evaluation not only recognizes that development, scaling, and innovation are manifestations of complex adaptive systems, but also that interventions and actors must adapt their designs and actions in response to the changing environment, and that the tools and techniques in an

[21] One of us, Walton, was a co-PI on one of the RCTs, that explored an intermediate scaling in two states, and is the co-author of the synthesis paper.

evaluation must also adapt to the nature of the task, the level of complexity, and the type of question.

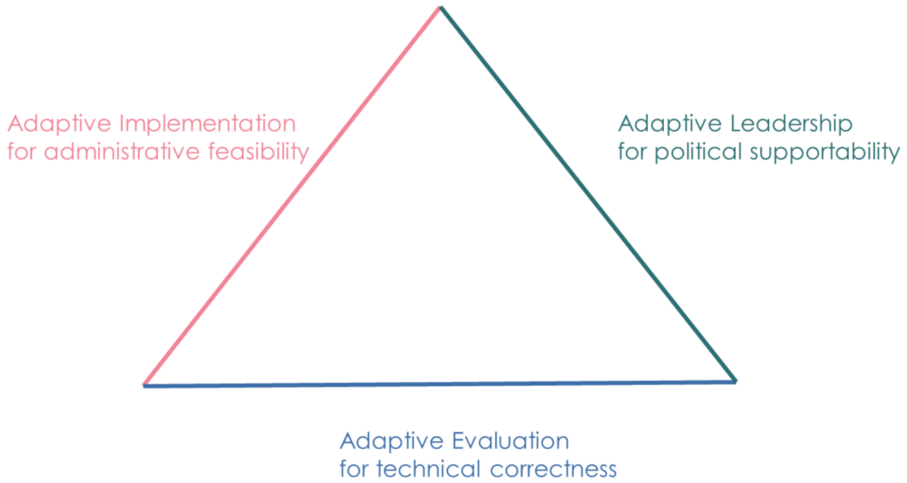
Adaptation cannot be assumed but comes with its own set of challenges. It involves an aligned mindset among multiple stakeholders. It requires both an authorizing environment for change, that recognizes the need and supports the exploration of alternatives, and actors who are capable and motivated to undertake the exploration process within the organization and broader system. It requires creating a safe space to fail and fail often, and the creation of mechanisms to then understand the failures and learn from them. The participatory nature of the evaluation, which is a bedrock of the Adaptive Evaluation philosophy, demands a high level of trust, communication and coordination between the evaluation, implementation teams, and broader community of stakeholders.

Adaptive evaluation can be seen as a powerful complement to two other elements of the change process that have not been formally elucidated in this paper—adaptive leadership and adaptive implementation. **Adaptive leadership** creates the enabling environment for meaningful change to take place (Heifetz et al., 2009). This involves creating the space in which unspoken tensions in a system can surface, in which stakeholders respect and empathize with one another until there is no *other* and they are *one* in the task of enacting change, and in which people closest to the problem are empowered to take ownership of their situation. It involves creating the space to fail. **Adaptive implementation** involves putting in place the structures and mechanisms for rapid prototyping, experimentation, and learning. This is at the centre of agile processes, and sequential design sprints in the private and public sectors; while a closely related approach for the public sector is the Problem Driven Iterative Adaptation (Andrews et al, 2017). An **Adaptive Evaluation** complements these two, ensuring that the evaluation respects the complexity of desired change, is nimble enough to adapt to changes in implementation and the environment, and is structured to provide feedback into the iterative design process. An Adaptive Evaluation, too, is challenging. A holistic view of complexity, that encompasses innovation and scaling requires expertise in a variety of tools. It requires discipline, to tackle multiple hypotheses and involve all relevant parties, but remain objective.

And thus, we have a trifecta—Adaptive Implementation, Adaptive Leadership, and Adaptive Evaluation. This aligns closely with the policy triangle developed at the Harvard Kennedy School, in which a policy must be administratively feasible (that, is implementable), politically supportable (there is an effective coalition and associated leadership) and technically correct (that there is a complete causal chain from the policy intervention to the outcome).[22] The adaptive trifecta is effectively an adaptive approach to address each part of the triangle. The desired change is materialized when these three elements are working together in tandem.

[22] This is adapted from Moore, 2013

Figure 35: The Adaptive Trifecta



This paper outlines the theoretical justification for an Adaptive Evaluation. In parallel, this is being supplemented with case studies and applications, from organizations that did this in the past, and from recent Adaptive Evaluations. Learning from doing will help refine the structure and design of an Adaptive Evaluation, which in the true spirit of adaptation, we expect will continue to evolve. Currently, we are personally involved in four Adaptive Evaluations running in India and Brazil. We hope more will emerge, and that there is more deep ongoing research that builds onto this paper by carefully documenting the challenges and benefits of this complexity-based approach.

6. Bibliography

Aiyar, Y., & Bhattacharya, S. (2016). The post office paradox: A case study of the block level education bureaucracy. *Economic and Political Weekly*, 61-69.

Amabile, T. M. (1988). A model of creativity and innovation in organizations. *Research in organizational behavior*, *10*(1), 123-167.

An, Z., Jalles, J. T., & Loungani, P. (2018). How well do economists forecast recessions?. *International Finance*, *21*(2), 100-121.

Anderson, R. A., Crabtree, B. F., Steele, D. J., & McDaniel Jr, R. R. (2005). Case study research: The view from complexity science. *Qualitative health research*, *15*(5), 669-685.

Andrews, M., Pritchett, L., & Woolcock, M. (2017). *Building state capability: Evidence, analysis, action* (p. 288). Oxford University Press.

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2), 3-30.

Arthur, W. B. (2009). Complexity and the economy. In *Handbook of Research on Complexity*. Edward Elgar Publishing.

Ball, P. (2001). *The self-made tapestry: pattern formation in nature*. Oxford University Press, Inc..

Bamberger, M., Vaessen, J., & Raimondo, E. (Eds.). (2015). *Dealing with complexity in development evaluation: A practical approach*. Sage Publications.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, *122*(3), 1235-1264.

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy, 2(*1), 1-30.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., ... & Walton, M. (2016). *Mainstreaming an effective intervention: Evidence from randomized evaluations of "Teaching at the Right Level" in India* (No. w22746). National Bureau of Economic Research.

Barder, O. (2012a, August 20). *The implications of complexity for development*. https://www.cgdev.org/media/implications-complexity-development-owen-barder

Barder, O. (2012b, August 16). What is development? *Center for Global Development | Ideas to Action*. https://www.cgdev.org/blog/what-development

Beinhocker, E. D. (2006). *The origin of wealth: Evolution, complexity, and the radical remaking of economics*. Harvard Business Press.

Bennett, A., & Checkel, J. T. (Eds.). (2015). *Process tracing*. Cambridge University Press.

Bhanjdeo, A., Narain, N., Sheth, S., & Walton, M. (2021). *Seeing like a State or Seeing the State? A Qualitative Study of a Government Program to Support Women's Self-Help Groups in Madhya Pradesh, India.* Imago Global Grassroots. https://imagogg.org/ideas-and-insights/seeing-like-a-state-or-seeing-the-state/

Biswas, H. R., Hasan, M. M., & Bala, S. K. (2018). Chaos theory and its applications in our real life. *Barishal University Journal Part*, *1*(5), 123-140.

Blackman, T., Wistow, J., & Byrne, D. (2013). Using qualitative comparative analysis to understand complex policy problems. *Evaluation*, *19*(2), 126-140.

Brockmann, D. (2018a, January 12). *"Double trouble" The double pendulum.* Complexity Explorables. https://www.complexity-explorables.org/explorables/double-trouble/

Brockmann, D. (2018b, January 28). *"Lotka martini" The Lotka-Volterra model.* Complexity Explorables. https://www.complexity-explorables.org/explorables/lotka-martini/

Brockmann, D. (2019, March 28). *"Berlin 8:00 a.m." The Emergence of Phantom Traffic Jams.* Complexity Explorables. https://www.complexity-explorables.org/explorables/berlin-8-am/

Byrne, D. (2005). Complexity, configurations and cases. *Theory, culture & society*, *22*(5), 95-111.

Byrne, D. (2013). Evaluating complex social interventions in a complex world. *Evaluation*, *19*(3), 217-228.

Byrne, D., & Callaghan, G. (2013). Complexity theory and the social sciences: The state of the art. Routledge.

Camazine, S., Deneubourg, J. L., Franks, N. R., Sneyd, J., Theraula, G., & Bonabeau, E. (2020). Self-organization in biological systems. In *Self-Organization in Biological Systems*. Princeton university press.

Carter, B. (2012). Helpdesk Research Report: Theory-based evaluation approach. *Birmingham: Governance and Social Development Resource Centre*.

Cartwright, J. (2013, March 8). Physicists Discover a Whopping 13 New Solutions to Three-Body Problem. *Science.* https://www.science.org/content/article/physicists-discover-whopping-13-new-solutions-three-body-problem

Cedric , S., & McIntyre , A. (2019, April 10). Economic Forecasting Is Really Difficult. Just Ask the IMF. *Bloomberg.* https://www.bloomberg.com/graphics/2019-imf-forecasts/

Collier, D. (2011). Understanding process tracing. *PS: Political Science & Politics*, *44*(4), 823-830.

Complex Adaptive Systems. (2015, May 4). Systems Innovation. https://www.youtube.com/watch?v=IWhkUne8T68

Cooley , L., & Guerrero, I. (2016). *The broken part of the business model in taking development outcomes to scale.* Imago Global Grassroots. Retrieved July 31, 2022, from https://imagogg.org/ideas-and-insights/the-broken-part-of-the-business-model-in-taking-development-outcomes-to-scale/

Coryn, C. L., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American journal of Evaluation*, *32*(2), 199-226.

Developmental evaluation. (2011, April 30). BetterEvaluation. https://www.betterevaluation.org/en/plan/approach/developmental_evaluation

De Domenico, M., & Sayama, H. (2019). *Complexity explained.* https://complexityexplained.github.io/

Deshpande, A. (2022). *Rural livelihoods programs in South Asia: Lessons from the Past as a Guide to the Future.* https://thedocs.worldbank.org/en/doc/62141abf1152d94953e76c11345c869f-0310022021/related/SARL-India-Final-Report-Final-May22.pdf

DeSilver, D. (2019). *Despite global concerns about democracy, more than half of countries are democratic.* Pew Research Center. https://www.pewresearch.org/fact-tank/2019/05/14/more-than-half-of-countries-are-democratic/

Dobzhansky, Theodosius. "On some fundamental concepts of Darwinian biology." *Evolutionary biology*. Springer, Boston, MA, 1968. 1-34.

Duflo, E., Berry, J., Mukerji, S., & Shotland, M. (2015). A wide angle view of learning.

El Houda, R., Gokhale, S., Lestani, D., Sidders, M., & Yang, Y. (2020). *Understanding Agricultural Productivity in Nigeria.*

Farr, J. L., & West, M. A. (Eds.). (1990). *Innovation and creativity at work: Psychological and organizational strategies*. Wiley.

Felipe, J., Usui, N., & Abdon, A. (2011). Rethinking the growth diagnostics approach: questions from the practitioners. *Journal of International Commerce, Economics and Policy*, *2*(02), 251-276.

Forrester, J. W. (1999). System dynamics: the foundation under systems thinking. *Sloan School of Management. Massachusetts Institute of Technology*, *10*.

Forss, K., Marra, M., & Schwartz, R. (Eds.). (2011). *Evaluating the complex: Attribution, contribution, and beyond* (Vol. 1). Transaction Publishers.

Freeman, L. (2006). The Development of Social Network Analysis. Vancouver: Empirical Pres; Wellman, Barry and SD Berkowitz, eds., 1988. Social Structures: A Network Approach.

Gallegati, Mauro, and Alan Kirman. "Reconstructing economics: Agent based models and complexity." *Complexity Economics* 1.1 (2012): 5-31.

Gamble, J. A. (2008). A developmental evaluation primer.

Gare, Arran. "Systems theory and complexity: introduction." (2000): 327-339.

GDP predictions are reliable only in the short term. (2018, December 15). *The Economist.* https://www.economist.com/graphic-detail/2018/12/15/gdp-predictions-are-reliable-only-in-the-short-term

Gokhale, S. (2020, May 14). Climate change means erratic monsoons, and erratic monsoons come at a high cost – the wire science. *The Wire Science.* https://science.thewire.in/environment/climate-change-india-rainfall-monsoons/

Gopal, S., & Clarke, T. (n.d.). *Guide to Actor Mapping. FSG. Retrieved July 9, 2022, from https://www.fsg.org/resource/guide-actor-mapping/*

Gleick, J. (2011). Chaos: Making a new science (Enhanced edition). *Open Road Media*, *1007*.

Glennerster, R. (2017, November 15). *When do innovation and evidence change lives?* The Abdul Latif Jameel Poverty Action Lab (J-PAL). https://www.povertyactionlab.org/blog/11-15-17/when-do-innovation-and-evidence-change-lives

Glennerster, R., & Takavarasha, K. (2013). Running randomized evaluations. In *Running Randomized Evaluations*. Princeton University Press.

Glouberman, Sholom, and Brenda Zimmerman. *Complicated and complex systems: what would successful reform of Medicare look like?*. Vol. 8. Toronto: Commission on the future of health care in Canada, 2002.

Gokhale, S. (2020, May 14). Climate change means erratic monsoons, and erratic monsoons come at a high cost – the wire science. *The Wire Science.* https://science.thewire.in/environment/climate-change-india-rainfall-monsoons/

Goldstein, J. (1999). Emergence as a construct: History and issues. *Emergence*, *1*(1), 49-72.

Gopal, S., & Clarke, T. (n.d.). *Guide to Actor Mapping. FSG. Retrieved July 9, 2022, from https://www.fsg.org/resource/guide-actor-mapping/*

Green, D. (2018, February 13). *10 things to know about adaptive experimental design*. EGAP. https://egap.org/resource/10-things-to-know-about-adaptive-experimental-design/

Guerrero, I., & Walton, M. (2015). *Why Is It so Hard for Organizations at the Base of the Pyramid to Go to Scale? The Missing Middle*. Imago Global Grassroots. https://imagogg.org/ideas-and-insights/the-missing-middle-why-is-it-so-hard-for-organizations-at-the-base-of-the-pyramid-to-go-to-scale/

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., & Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, *118*(15), e2014602118.

Harford, T. (2011). *Adapt: Why success always starts with failure*. Farrar, Straus and Giroux.

Hari, S., Khare, P., & Subramanian, A. (2018, June 8). *Climate change and Indian agriculture.* VoxDev. https://voxdev.org/topic/agriculture/climate-change-and-indian-agriculture

Hausmann, R., Rodrik, D., & Velasco, A. (2008). Growth diagnostics. *The Washington consensus reconsidered: Towards a new global governance*, *2008*, 324-355.

Heifetz, R. A., Heifetz, R., Grashow, A., & Linsky, M. (2009). *The practice of adaptive leadership: Tools and tactics for changing your organization and the world.* Harvard Business Press.

Holland, J. H. (1992a). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

Holland, J. H. (1992b). Complex adaptive systems. *Daedalus*, *121*(1), 17-30.

Hummelbrunner, R. (2000). A Systems Approach to Evaluation: Application of Systems Theory and Systems Thinking in Evaluation. Paper prepared for the 4 th EEC Conference, Lausanne, Switzerland.

Implementation of NRLM. (2022, July 24). About DAY-NRLM; National Rural Livelihoods Mission (NRLM), Ministry of Rural Development - Govt. of India. *https://aajeevika.gov.in/en/content/implementation*

In India, Women's Self-Help Groups combat the COVID-19 (Coronavirus) pandemic. (2020, April 11). *World Bank.* https://www.worldbank.org/en/news/feature/2020/04/11/women-self-help-groups-combat-covid19-coronavirus-pandemic-india

Ishimatsu, H., Sugasawa, Y., & Sakurai, K. (2004). Understanding innovation as a complex adaptive system: Case studies from Shimadzu and NEC. *Pacific Economic Review*, *9*(4), 371-376.

ITRAC. (2017). *Theory Based Evaluation*. ITRAC for Civil Society. https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Theory-based-evaluation.pdf

Kast, F. E., & Rosenzweig, J. E. (1972). General systems theory: Applications for organization and management. *Academy of management journal*, *15*(4), 447-465.

Kasy, M., & Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, *89*(1), 113-132.

Krugman, P. (1996). *The self organizing economy*. John Wiley & Sons.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1), 31-43.

Levy, S. (2019, February 14). Santiago Levy on the Scaling Up of Conditional Cash Transfers in Mexico [Interview]. https://www.youtube.com/watch?v=MhuAXY1-YQo

Lorenz, E. (1972). *Predictability: does the flap of a butterfly's wing in Brazil set off a tornado in Texas?* (p. 181). na..

Mitchell, M. (2009). *Complexity: A guided tour*. Oxford university press.

Moore, M. H. (2013). *Recognizing public value*. Harvard University Press.

Morçöl, G. (2013). *A complexity theory for public policy*. Routledge.

Myrdal, G. (1974). What is development?. Journal of Economic Issues, 8(4), 729-736.

Narayanan, S., Oldiges, C., & Saha, S. (2020). *Employment guarantee during times of COVID-19: Pro-poor and pro-return-migrant?* (No. 2020-034). Indira Gandhi Institute of Development Research, Mumbai, India.

Pagels, H. R. (1989). *The dreams of reason: the computer and the rise of the sciences of complexity*. Toronto ; New York: Bantam Books.

Paparini, S., Papoutsi, C., Murdoch, J., Green, J., Petticrew, M., Greenhalgh, T., & Shaw, S. E. (2021). Evaluating complex interventions in context: systematic, meta-narrative review of case study approaches. *BMC medical research methodology*, 21(1), 1-22.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. Guilford press.

Pawson, R., Tilley, N., & Tilley, N. (1997). *Realistic evaluation*. sage.

Pines, D. (2014). *Emergence: A unifying theme for 21st century science. Santa Fe Institute. https://medium.com/sfi-30-foundations-frontiers/emergence-a-unifying-theme-for-21st-century-science-4324ac0f951e*

Pritchett, L. (2009). Is India a flailing state?: Detours on the four lane highway to modernization.

Raimondo, E. (2020). Getting Practical With Causal Mechanisms: The application of Process-Tracing Under Real-World Evaluation Constraints. *New Directions for Evaluation*, *2020*(167), 45-58.

Ratan, A. L. (2022). *Scaling convergence for social and economic services & programmes in rural India: Conceptual framework & evidence review*. Imago Global Grassroots. https://imagogg.org/ideas-and-insights/convergence/

Realist Evaluation. (2014, September 10). BetterEvaluation. https://www.betterevaluation.org/en/approach/realist_evaluation

Rodrik, D. (2006). Goodbye Washington consensus, hello Washington confusion? A review of the World Bank's economic growth in the 1990s: learning from a decade of reform. *Journal of Economic literature*, *44*(4), 973-987.

Raimondo, E. (2020). Getting Practical With Causal Mechanisms: The application of Process-Tracing Under Real-World Evaluation Constraints. *New Directions for Evaluation*, *2020*(167), 45-58. Rodrik, D. (2015). *Economics rules: The rights and wrongs of the dismal science*. WW Norton & Company..

Room, G. (2011). *Complexity, institutions and public policy: Agile decision-making in a turbulent world*. Edward Elgar Publishing.

Rodrik, D. (2015). *Economics rules: The rights and wrongs of the dismal science*. WW Norton & Company..

Schwartz, J. (2021, June 4). A Million Years of Data Confirms: Monsoons Are Likely to Get Worse. *The New York Times*. https://www.nytimes.com/2021/06/04/climate/monsoons-climate-change.html

Sen, A. (1999). Commodities and capabilities. *OUP Catalogue*

Sen, A. (2014). Development as freedom (1999). *The globalization and development reader: Perspectives on development and global change*, *525*.

Snowden, D. J., & Boone, M. E. (2007). A leader's framework for decision making. *Harvard business review*, *85*(11), 68.

Stame, N. (2004). Theory-based evaluation and types of complexity. *Evaluation*, *10*(1), 58-76.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). Broadening the range of designs and methods for impact evaluations.

Stiglitz, J., & Schoenfelder, L. (2003). Challenging the Washington consensus. *The Brown Journal of World Affairs*, 9(2), 33-40.

Strathern, M., & McGlade, J. (Eds.). (2014). *The Social Face of Complexity Science: A Festschrift for Professor Peter M. Allen*. Emergent Publications.

Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press.

Tao, Y. (2016). Spontaneous economic order. *Journal of Evolutionary Economics*, *26*(3), 467-500.

Thompson, WR (1933), "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples", Biometrika, 25(3/4): 285–294.

Turner, J. R., & Baker, R. M. (2019). Complexity theory: An overview with potential applications for the social sciences. *Systems*, *7*(1), 4.

Turner, J. R., & Baker, R. (2020). Just doing the do: A case study testing creativity and innovative processes as complex adaptive systems. *New Horizons in Adult Education and Human Resource Development*, *32*(2), 40-61.

Unger, R. M. (2000). *Democracy realized: The progressive alternative*. Verso.

Unger, R. M. (2014, January 9). *What is wrong with the social sciences today?* [Interview]. https://www.socialsciencespace.com/2014/01/roberto-mangabeira-unger-what-is-wrong-with-the-social-sciences-today/

Von Bertalanffy, L. (1972). The history and status of general systems theory. *Academy of management journal*, *15*(4), 407-426.

Waddell, S. (2016). *Change for the Audacious: A Doer's Guide*. NetworkingAction Publishing.

Walton, M. (2014). Applying complexity theory: A review to inform evaluation design. *Evaluation and Program Planning*, *45*, 119-126.

Widner, J., Woolcock, M., & Nieto, D. O. (2022). The Case for Case Studies.

Williamson, J. (1990). Latin American adjustment: How much has happened?.

Wilson-Grau, R., & Britt, H. (2012). Outcome harvesting. *Cairo: Ford Foundation*.

Yilmaz, L. (2008). Innovation Systems Are Self-organizing Complex Adaptive Systems. In *AAAI Spring Symposium: Creative Intelligent Systems* (pp. 142-148).

Zagha, R., & Nankani, G. T. (Eds.). (2005). Economic Growth in the 1990s: Learning from a Decade of Reform. World Bank Publications.

Zimmerman, B. J., Dubois, N., Houle, J., Lloyd, S., Mercier, C., Brousselle, A., & Rey, L. (2012). How does complexity impact evaluation? An introduction to the special issue. *The Canadian journal of program evaluation= La Revue canadienne d'evaluation de programme*, *26*(3), v.

FIND YOUR VOICE.
SPREAD YOUR WINGS
REIMAGINE THE FUTURE.

IMAGO
Global Grassroots