Centre of Excellence for Development Impact and Learning

DECEMBER 2021

CEDIL Methods Working Paper 6

Evaluating complex interventions in international development

Edoardo Masset, Som Shrestha and Matt Juden

LIDC London International Development Centre

UKaid from the British people

Oxford Policy Management

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL methods working papers

The CEDIL methods working paper series offers innovative research methods for developing impact evaluation and evidence synthesis work in low- and middle-income countries.

About this working paper

Please direct any comments or queries to the corresponding author, Edoardo Masset at Edoardo.Masset@lidc.ac.uk

Suggested citation: Masset, E., Shrestha, S. and Juden, M. (2021). 'Evaluating Complex Interventions in International Development'. *CEDIL Methods Working Paper 6*. Centre of Excellence for Development Impact and Learning (CEDIL), London and Oxford [online]. Available from: https://doi.org/10.51744/CMWP6

This project was funded by the Centre of Excellence for Development Impact and Learning (CEDIL), supported by UK Aid from the UK Government. The views expressed in this working paper do not necessarily reflect the UK Government's official policies.

Cover design: PhilDoesDesign

CEDIL methods working paper: Evaluating complex interventions in international development

Authors

Edoardo Masset
London School of Hygiene and Tropical Medicine

Som Shrestha
London School of Hygiene and Tropical Medicine

Matt Juden
London School of Hygiene and Tropical Medicine

Table of contents

List of figures

List of tables

List of boxes

List of abbreviations

BRAC	Bangladesh Rural Advancement Committee
CECAN	Centre for the Evaluation of Complexity Across the Nexus
CGE	Computable General Equilibrium
COMPASSS	Comparative Methods for Systematic Cross-Case Analysis
DFID	UK Department for International Development
ICAI	Independent Commission for Aid Impact
JASSS	Journal of Artificial Societies and Social Simulation
JCEOP	Jordan Compact Economic Opportunities Programme
M4P	Making Markets Work for the Poor
MRC	Medical Research Council
NGO	Non-Governmental Organisation
PMI	President's Malaria Initiative
QCA	Qualitative Comparative Analysis
RCT	Randomised Control Trial
SUTVA	Stable Unit of Treatment Value Assumption
UPGP	Ultra Poor Graduation Programme
USAID	United States Agency for International Development

1. Introduction

The aim of this paper is to provide guidance on the evaluation of complex interventions in international development. Our contribution to the literature is threefold. First, unlike other reviews on the same subject, our focus is exclusively directed to evaluations of development interventions. Second, we identify methods that are new or little used in the evaluation literature. Third, we map methods to project types with the aim of identifying methods that are useful to the evaluation of particular projects.

We start by making a distinction between complex interventions and complex systems. While much recent research has been devoted to understanding how interventions can be implemented and understood in complex systems, our interest rests primarily with the goal of understanding and evaluating complex interventions. We therefore abstract from complexities arising from the interaction between interventions and complex systems. In our framework, complex interventions can be implemented in simple as well as in complex systems, but the latter are not the primary focus of our paper. Removing the consideration of interactions between interventions and complex systems reduces the scope of our review, but helps focus attention on a manageable number of issues and methods.

We define complex interventions as interventions characterised by multiple components, multiple stakeholders, or multiple target populations, or interventions incorporating multiple processes of behavioural change. These interventions cannot be represented by single-intervention or single-outcome models, and present several challenges to evaluation. Occasionally, or purposely, these interventions give rise to emergent outcomes such as non-linear effects, tipping points, and multiple equilibria. The presence of emergent outcomes of this type adds additional difficulties to evaluation because existing methods are ill-equipped to detect and estimate non-linear impacts of interventions.

Complex interventions are extremely common in international development. In fact, most development interventions display some element of complexity, as defined above, and it is probably more correct to speak of degrees of complexity than to build classifications of ‘simple’ and ‘complex’, or ‘complicated’, interventions. The evaluation of complex interventions presents various challenges. Researchers normally neglect complexity and evaluate interventions as though they were not complex, for example singling out the impact of a single component or by giving up on the evaluation of complex interventions altogether. One of the reasons for this is that methods for the

evaluation of complex interventions are somewhat underexplored or poorly understood.

Our paper reviews promising methods for the evaluation of complex interventions that are new or little used. Few of the methods reviewed are new, but their application to evaluation or to the evaluation of development interventions is. The empirical application of these methods to the evaluation of development interventions has been very limited (or nil). Indeed, a greater application of these methods would require much testing and refinement. With few exceptions, our review shows that researchers are evaluating complex development interventions using a small number of methods that are not fit for purpose.

In our review, we restrict our attention to methods that address causality, and therefore allow us to state conclusively whether an intervention works or not. These include counterfactual methods in a broad sense. We include methods that compare changes in outcomes in an intervention group against changes that would have occurred in the absence of the intervention. The counterfactual may consist of an experimental control group, of a multiplicity of groups receiving different interventions, or of a hypothetical intervention. This definition allows the inclusion of methods exploring multiple causes of effects, such as factorial designs and qualitative comparative analysis (QCA), and of methods using simulations of interventions, such as structural modelling and agent-based modelling. Qualitative and quantitative approaches that do not carry out an explicit or implicit comparison between the intervention and what would have happened in the absence of the intervention are not covered in our review.

The paper is structured in the following way. Section 2 reviews the existing literature on the evaluation of complex interventions and presents the original contribution of our work. Section 3 illustrates a taxonomy of complex interventions in international development. Section 4 discusses the fundamental challenges faced by researchers evaluating complex interventions, and the evaluation methods they most commonly employ. Section 5 discusses a series of promising impact evaluation methods for the evaluation of different types of complex interventions, drawing on a review of the literature and a large number of examples. Section 6 discusses the limitations of our review. Section 7 summarises the main results of the reviews and concludes.

2. Existing guidelines to the evaluation of complex interventions

There is a large body of literature, both theoretical and empirical, on the evaluation of complex interventions. In this section, we review this literature by summarising the contents of three guidelines documents: the guidelines on developing and evaluating complex interventions issued by the Medical Research Council (MRC) (Craig *et al.*, 2019); the UK Government's *Magenta Book 2020 Supplementary Guide: Handling Complexity in Policy Evaluation* (HM Treasury, 2020); and *Dealing with Complexity in Development Evaluation* (Bamberger, Vaessen and Raimondo, 2016), edited by the Independent Evaluation Group of the World Bank.

These three sets of guidelines were written by teams with different epistemological perspectives. The MRC team consists of epidemiologists and statisticians drawing on the experimental tradition to evaluation, although the latest version of the MRC guidelines includes realist and system approaches. The Supplement to the Magenta Book was compiled by the Centre for the Evaluation of Complexity Across the Nexus (CECAN) team, which comprises social scientists and policy analysts sharing an interest in complexity science, system approaches, and other methodologies, such as Bayesian and configurational analysis. The book edited by the World Bank was developed by a team of economists and evaluators with practical experience in the use of quasi-experimental methods, qualitative methods, and mixed methods. Taken together, the three guidelines therefore present a comprehensive survey of the methods that have been proposed in the literature for the evaluation of complex interventions.

In this section, we describe how the different guidelines define complex interventions, the main challenges to evaluation they identify, and the approaches and methods they propose. We then discuss the original contribution made by our review: our focus on complex interventions rather than on complex systems, our exclusive attention on impact evaluation methods that address causality, and our attempt to map evaluation methods to specific types of project.

2.1 The MRC guidelines on developing and evaluating complex interventions

Over the last 20 years, the MRC has published (or nearly published, the last being only a draft) three sets of guidelines on the design and evaluation of complex interventions. The first guidelines (MRC, 2000), drawing on an earlier paper by Campbell *et al.* (2000), defined complex interventions as interventions comprising a number of components,

whose 'active ingredient' was difficult to specify. The guidelines recommended a 'step-wise' procedure for randomised trials of complex interventions. In addition, they recommended that trials should be preceded by a theoretical understanding of how interventions work, by modelling how different components operate and interact, and by piloting the effectiveness of single-intervention components in small trials. The guidelines mentioned in passing the use of qualitative methods to understand the operation of mechanisms, but the method of choice remained the randomised control trial (RCT).

The revised MRC guidelines (Craig *et al.*, 2019) published in 2006 recognised that all interventions are, to some extent, complex and that complexity is not dichotomous but more a matter of degree. The new guidelines defined complex interventions more simply as consisting of multiple interacting components. More precisely, they identified the following dimensions of complexity: the number of interactions between components; the number and difficulty of behaviours required for the intervention to be successful; the number of groups and organisations targeted by the intervention; the number and variability of outcomes; and the degree of adaptation of the intervention.

These revised guidelines restated the difficulty of identifying the 'active ingredient' of complex interventions and stressed the sensitivity of the estimated impacts to the local context, and the difficulty of employing experimental designs. They recommended using the evaluation design that is most appropriate in each particular case. The method of preference was still the RCT, particularly in the form of the cluster randomised trial, stepped-wedge designs, preference trials, and N-of-1 designs. However, it was recognised that when experiments are not possible, researchers should use quasi-experimental methods such as regression analysis, instrumental variables, and matching methods. They also recommended conducting a process evaluation in all cases, in order to better understand how an intervention works. The revised guidance was later complemented by more detailed guidance on process evaluation (Moore *et al.*, 2015) and on quasi-experimental designs (Craig *et al.*, 2019).

The draft of the new revised MRC guidelines (Skivington *et al.*, 2021) marks a significant shift in the definition of complex interventions. Complex interventions are no longer defined by the number of components and other characteristics, which are now considered as dimensions or degrees of 'complicatedness'. Complexity is defined as a property of the system in which the intervention operates, with complex systems being those exhibiting properties such as emergence, feedback loops, adaptation, and self-organisation. The latest guidelines present four different approaches to the evaluation of complex interventions: efficacy, effectiveness, realist, and system approaches. The

approaches are recommended in order of system complexity for example, system approaches should be used when complexity is the highest. The inclusion of realist and system approaches expands the set of recommended methods. In addition to RCTs, quasi-experimental methods, and process evaluations, the new guidelines recommend the use of agent-based modelling, system dynamic modelling, network analysis, QCA, and realist trials (although the existence of 'realist trials' has been questioned by some: see in particular the debate in Van Belle *et al*., 2016 and Bonell *et al.*, 2016).

2.2 Magenta Book Supplementary Guide: Handling Complexity in Policy Evaluation

The Magenta Book is the UK Government official guidance on policy evaluation. *Magenta Book Supplementary Guide: Handling Complexity in Policy Evaluation* was prepared by CECAN to inform evaluations of interventions in complex environments. The Supplement often refers to 'complex interventions' and 'complex evaluations', although these are not precisely defined. Characteristics of complex interventions are discussed among the 'drivers' of complexity, which include policies delivered at different levels and involving different interventions; interventions that involve many actors and that generate conflicts of perspectives; interventions addressing more than one policy domain; and cases in which no single organisation has full control over the intervention and its outcomes. The Supplement defines the characteristics of complex systems in great detail. Complex systems are composed of many components interacting in non-linear ways and producing phenomena discussed in complexity science, such as adaptation, emergence and self-organisation, unexpected indirect effects, feedback loops, levers and hubs, non-linearity, domains of stability, tipping points, path dependence, openness, and change over time. The Supplement states that understanding the characteristics of complex systems is key to understanding and improving public policies.

The Magenta Book Supplement identifies the following challenges in evaluating interventions in complex environments. First, context matters: the impacts of the interventions will differ depending on the characteristics of the context and of the range of intermediate and final outcomes considered. Second, standardisation of interventions in complex environments is difficult and even not desirable, as interventions need to adapt to the characteristics of the environment. In these circumstances, researchers are unable to 'control' for other factors, as is normally the case with randomised experiments and quasi-experimental designs. Third and finally, complex systems and complex interventions change over time, making prospective

evaluations unfeasible. Because of non-linear responses, the outcomes of activities are hard to predict, and prospective evaluation designs are not possible.

The Supplement concludes by describing six approaches (and related methods) that are well suited to evaluate policies in complex systems:

1. participatory approaches, such as developmental evaluation, action research, and peer challenge;
2. theory-based approaches, such as system mapping, realist evaluation, and contribution analysis;
3. configurational approaches, such as QCA;
4. counterfactual approaches, such as experiments, quasi-experiments, and simulation methods;
5. statistical approaches, such as regression analysis; and
6. synthesis designs, such as realist synthesis.

The guidelines do not recommend any approach in particular; they recommend that any choice of method should be based on the specific characteristics of the complex system, the purpose of the evaluation, and the feasibility of the approach. However, when the system is very complex (displaying multiple interactions, non-linearities, multiple causality, and so forth), they argue against the use of experiments and quasi-experiments. Instead, they recommend the use of a wide range of methods: generative causation (realist evaluation and contribution analysis), system mapping, system modelling, predictive modelling, scenario analysis, big data and associated methods, adaptive methods, QCA, participatory approaches, narrative methods, and agent-based modelling.

2.3 Dealing With Complexity in Development Evaluation: A Practical Approach by the Independent Evaluation Group of the World Bank

The Independent Evaluation Group of the World Bank edited a set of contributions on different topics relating to the evaluation of complex interventions. As an edited book, it does not contain a coherent set of guidelines. However, the introductory chapters provide definitions of complex interventions and complex systems, a description of the main challenges to evaluation, and a set of broad recommendations on best practices and approaches.

The book does not precisely define complex systems, but defines 'complexity' as characterised by unintended outcomes and multiple, non-linear, and recursive pathways. The authors argue that development interventions have become increasingly

complex over the last decade, and define complexity of interventions along the following five dimensions:

1. the nature of the intervention—scale and geographic spread (for example, cross-country interventions are more complex); diversity of the target population (programmes trying to address the different needs of heterogeneous populations); number and diversity of programme components with their interactions; organisational complexity (programmes including various implementers and stakeholders); and clarity and coherence of the objectives and of the theory of change;
2. institutions and stakeholders—the number and diversity of stakeholders; the number and diversity of funding sources; stakeholders' expectations and their level of consensus; and conflict or cooperation among implementing agencies;
3. causality of effects—attribution is difficult because the intervention affects institutions not just individuals; impacts appear in the long term rather than in the short term; and impacts are non-linear and difficult to capture;
4. nature of the system—the complexity of the political, social, and economic system in which the intervention is embedded; and
5. the complexity of the evaluation process—conducting evaluation in such contexts is a complex activity on its own.

The book advocates for complexity-responsive evaluations. Evaluations can become complexity-responsive by becoming aware of the large system in which interventions operate, by understanding how the various element of complexity affect the evaluation, and by explicitly addressing some of the complexity dimensions in the evaluation. The authors delineate four main strategies (and related methods) for complexity-responsive evaluations:

1. system thinking strategies to map out the complexity dimensions, including methods such as system mapping, social network analysis, agent-based modelling, system dynamics, and critical systems heuristics;
2. strengthening existing evaluation methodologies (including experimental, statistical, theory-based, case-based, and participatory approaches) to address complexity issues or using novel data for the analysis;
3. unpacking systems and reassembling them through approaches such as system modelling, statistical methods, portfolio analysis, comparative case analysis, and synthesis; and
4. holistic strategies seeking a comprehensive understanding of all the complexity dimensions of a particular intervention through the use of mixed methods.

2.4 Complexity interventions and complex systems

Discussions of complex interventions, in the guidelines and elsewhere, often make reference to 'complexity science'. However, it is not always clear what specific area of

methodological or theoretical enquiry researchers are referring to (Mowles, 2014). Complexity science is not a unified discipline and includes a wide variety of approaches (see, for example, the survey in Mitchell, 2009). In addition, complexity science has been primarily developed within mathematics, physics, and biology, and it is not obvious that its methods can be easily transported to the social sciences. In our review we will keep references to complexity science to a minimum.

We start here by defining complex interventions. A three-part distinction of projects into the categories of 'simple', 'complicated', and 'complex' has become popular in the evaluation literature. The distinction goes back to Glouberman and Zimmerman (2002), who classified social problems as simple (for example, following a recipe), complicated (for example, flying a rocket), and complex (for example, raising a child). Rogers (2008) employed this classification as an 'orienting framework for evaluation'; she defined as complex those projects characterised by feedback loops and tipping points and displaying emergent properties, and defined as complicated those interventions consisting of multiple components, multi-sites, or multi-governance elements. Many researchers have found this classification useful. However, we believe this classification produces an unnecessary simplification in the variety of development interventions, and we will not use it in our review. Complexity is a matter of degree rather than a dichotomous characteristic, and it is difficult to understand why, for example, 'flying a rocket' is just 'complicated' and does not include elements of complexity.

In our paper, we define as 'complex' a project made up of many components that interact in non-trivial ways. The interactions bring about outcomes that are more than the sum of the outcomes produced by the single components. In such projects, the outcomes of the whole intervention cannot be inferred from an understanding of the behaviour of each component separately considered.

There are several reasons why development interventions are complex as we have defined them. In some cases, interventions include many activities in the belief that the different components will produce 'complementarities' and 'synergies', whereby the impact of the interventions implemented together is larger than the impacts of the interventions independently implemented. In other cases, the implementers have limited knowledge of the effectiveness of the interventions and target different groups with different activities and at different levels in the hope that they will produce the desired effects, either independently or in combination. In yet other cases, complex interventions are the result of a political compromise where different social groups demand different interventions or in which different suppliers promote different

interventions and no single organisation has full control over the intervention or its outcomes.

Some authors have argued that project complexity is a characteristics of the system, not of the intervention (Shiell, Hawe and Gold, 2008; Greenhalgh and Papoutsi, 2018). A complex system consists of many components that interact in non-obvious ways and that generate emergent outcomes. We agree that any project, either simple or complex, may generate a variety of unexpected outcomes when implemented in a complex system. However, in our review, we will limit our analysis to complex projects and we will abstract from the characteristics of the system.

There are some limitations in adopting this approach. Abstracting from the complexity of the system reduces the number of project instances considered and the severity of the evaluation problems analysed. Many of the complex phenomena observed are the result of interactions between interventions and complex systems. These interactions generate two main issues. First, when the interventions interact with fixed characteristics of the society and of the economy, such as the geography of the demographic structure of the population, the effects of interventions are difficult to predict. Effects will occur depending on the presence of supporting and derailing factors in the specific context considered. In this case, the complexity of the system generates a problem of external validity because the interactions with the varied characteristics of the system give rise to different outcomes. Second, when the interventions interact with variable characteristics of the system, for example other simultaneously implemented interventions or government policies, further synergies between the project activities and the system may occur. In this case, it becomes increasingly difficult to understand how the project operates and to disentangle the impact of single project components.

By ignoring the complexity of the system, we abstract from problems of external validity and from complex interactions between the interventions and its environment. We are not making this choice in the belief that external validity and project-system interactions are not important. The choice is pragmatic, in that it allows us to focus on project complexity alone and to investigate some specific project types and evaluation approaches in greater detail.

2.5 Complex interventions and causality

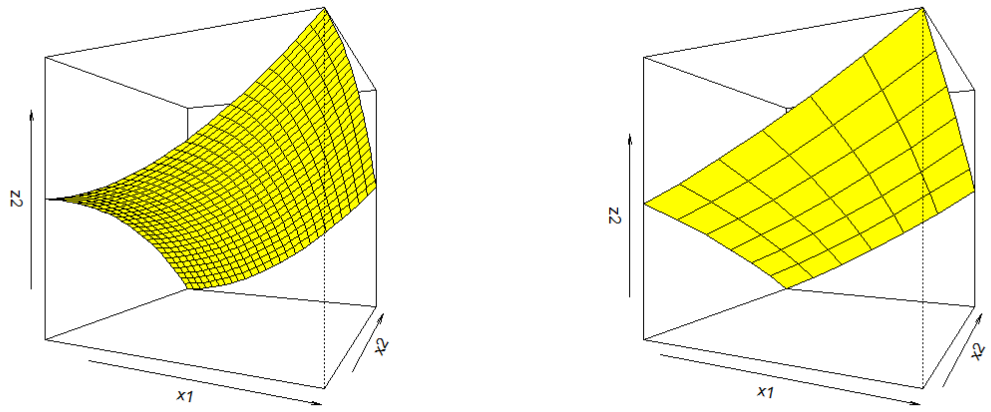
Our review focuses exclusively on methods that address causality. Some authors have claimed that methods of causal inference, such as experiments and quasi-experiments, cannot be employed in the evaluation of complex interventions. For example, Rogers (2009) argues that standard methods of causal inference can be employed in the

evaluation of 'complicated' interventions, but not in the evaluation of 'complex' ones. Hummelbrunner (2011) maintains that complex interventions cannot be evaluated by standard methods because the relationships between cause and effect can only be understood in retrospect, and because they are heavily dependent on initial conditions. Some have gone as far as to suggest that causality, as it is understood in experimental evaluation, cannot be identified at all in complex systems (Wagner, 1999). According to this view, non-linear interactions between variables result in emergent phenomena such as state transitions, limit cycles, and chaotic dynamics, which are totally unpredictable.

Views on the unfeasibility of using methods of causal inference in the evaluation of complex interventions appear exaggerated. They also pose a risk of 'complexification' of interventions, whereby programme managers or researchers can claim an intervention is complex as an excuse for not conducting a rigorous evaluation. There is scope for using standard methods of causal inference in the evaluation of complex interventions for at least two reasons. The first is that linear relationships between variable might be more common than it is thought, allowing the use of experiments and quasi-experiments. The relevance of non-linearities should be demonstrated with the data and observation, rather than being postulated.

Second, even if relationships are non-linear, this does not mean they cannot be analysed using linear methods. Non-linear relationships can be linear within the restricted range that is of policy interest. For example, the chart on the left-hand-side of Figure 1 represents a highly non-linear relationship. The response of the outcome variable *z* to changes in inputs *x1* and *x2* depends on the baseline values of the variables, and on the changes occurring in the other inputs. However, if we zoom in on the figure by restricting the range of the input variables, the response of the outcome becomes linear, as in the chart on the right-hand-side of Figure 1. Development interventions do not produce large changes in the inputs and normally operate in the short term, so that the linear model on the right-hand-side of Figure 1 might be the right one.

Figure 1: Linear approximation of a highly non-linear response



Although causal inference methods can be used to evaluate complex interventions, it must be recognised that the challenges in doing so are real. We summarise here the most common challenges faced by experimental and quasi-experimental designs. First, complex interventions are implemented at such a large scale that a valid control group cannot be identified. Second, multicomponent interventions can be evaluated experimentally with multi-arm trials, but the cost is prohibitive. Third, in multicomponent interventions, controlling for other activities and characteristics of the environment is of little use because the goal of the evaluation is precisely to understand how interventions affect outcomes under different circumstances. Fourth, complex interventions change over time in response to effects, or lack of effects, in such a way that protocols and prospective designs are not feasible.

These challenges are real and suggest that, although causality can be identified in complex interventions, this cannot be easily achieved using standard methods of causal inference, such as experiments and quasi-experiments. These methods need to be modified or combined with other methods in order to be able to address impacts that occur on a large scale, impacts that interact with multiple factors, and impacts that change over time. In our review, we will focus primarily on such modifications and combinations of methods. This choice does not imply that methods of causal inference can be applied in every context, or that they should be applied on their own. In general, other methods of enquiry should be used at the same time to better understand how projects work, and many of these methods will be qualitative.

2.6 Mapping complex interventions and evaluation methods

All the guideline documents illustrated above propose a wide range of methods to evaluate complex interventions, including experimental approaches, system-based methods, generative approaches, and qualitative methods, and encompassing nearly all quantitative and qualitative methods that are currently used in the social sciences. The reader in search for recommendations on how to evaluate specific interventions may be disappointed.

The guidelines also provide brief summaries of the recommended methods, and in some cases they provide examples of applications to real evaluation problems. However, they rarely recommend specific methods for a specific project. More often, they recommend that the researcher should adopt the method they feel is most appropriate to the specific context. At best, they provide general strategies for guiding the choice. For example, all guidelines seem to concur that, as contexts become more complex, researchers should rely increasingly more on generative and system-based approaches, and that evaluations should include combinations of different methods to address various evaluation challenges.

Our review follows a different approach. First, as mentioned, we restrict our analysis to methods that address causal inference. This greatly reduces the number of methods reviewed, because many of the approaches suggested for the evaluation of complex interventions are qualitative in nature. Whether a method addresses causal inference is not always easy to determine. In most cases, methods only address causal inference effectively under some conditions and assumptions. In our review, we decided to err on the side of caution by excluding methods that require strong assumptions for being causal, such as process tracing. Our decision is also pragmatic and dictated by the need to cover a small number of methods in greater detail, rather than covering a large number of them superficially.

The second difference with the other guideline documents is that we put more emphasis on identifying the most appropriate methods for specific interventions. Unlike other reviews, we first define typologies of complex interventions and indicate the main evaluation challenges they present. We then identify some promising methods to address these challenges and describe these methods in more detail through a number of applications found in the literature. Compared to other reviews, our task is made easier by our narrower focus on methods of causal inference. Our ultimate goal is to map out methods to specific complex interventions in order to offer quick guidance to researchers setting out to evaluate such interventions.

3. Types and characteristics of complex interventions

In this section, we provide a tentative taxonomy of complex interventions. We define 'typical' complex development interventions based on characteristics such as the number of activities involved, the number of targeted groups, and the scale of the intervention. We define the following project types: multicomponent interventions, portfolio interventions, long causal chain interventions, and system-level interventions. For each project type, we also provide examples from the practice of development.

This taxonomy is not comprehensive and does not cover all the complexities of development interventions. Not all projects will fall neatly into one of the typologies described. Indeed, many projects will share characteristics with more than one project type. In particular, characteristics described under the 'long causal chain' type will be recognisable in most development projects. Still, the typology will help us identify key challenges of complex evaluations and evaluation methods that fit specific project characteristics.

Our typology also abstracts from the characteristic of the context in which the interventions are implemented. Context matters, and some sectors of interventions (such as governance), are characterised by large contextual variability and by a multiplicity of actors implementing other policies and interventions. As mentioned in Section 2.4, this variability poses issues of external validity and limits the ability to attribute effects to interventions. By abstracting from the characteristics of the environment, we are allowed to ignore some aspects of evaluating complex interventions. We made this choice pragmatically, with the goal of focusing more in detail on the characteristics of the interventions.

3.1 Multicomponent interventions

Most development interventions include multiple components, and very few development projects consist of a single activity. Multicomponent interventions include several activities and are often implemented at the local level, for example a community or region. Activities can be carried out within a sector or can cover several sectoral areas, such as education, health, and livelihoods. Typically, these interventions expect to produce synergies. Project designers include different activities with outputs that interact and mutually reinforce each other in order to produce an impact on the project outcome that it is larger than the sum of the impacts produced by each activity separately implemented. Here, we further disaggregate multiple component

interventions into three types: 'separable interventions', 'synergistic projects', and 'critical mixes'. The distinction between these three types of intervention is not always neat, and many projects will contain combinations of them.

Separable interventions consist of different components targeting different populations or different behaviours. They are simultaneously implemented because the implementers hope to achieve higher population-level impacts by reaching different groups at the same time. In some cases, the implementers have limited knowledge of what is effective in addressing a specific problem. This is particularly true when there is limited generalisable knowledge available, such as knowledge from systematic reviews of evidence or other highly reliable sources, and when there is limited understanding of the context in which the project is implemented. In these circumstances stakeholders adopt several initiatives in the hope that some will be successful (Oliver *et al.*, 2021). They are not expecting the activities to interact and produce synergistic effects, although they may produce such effects. Their interest is to target different groups, or the same group, with different activities in the hope that some will be successful or that some unpredictable synergies may emerge.

For example, an intervention to control sexually transmitted infections (García *et al.*, 2012) included four activities: strengthening syndromic management by clinicians; mobile-team outreach for female sex workers; periodic presumptive treatment of trichomoniasis among female sex workers; and condom promotion for the general population. In another example (Walque *et al.*, 2017), a performance-based financing intervention provided clinics with financial incentives linked to results, additional resources available at the point of service delivery (not linked to performance), and enhanced supervision, coaching, and monitoring. In both examples, the different activities aimed to tackle the same problem in different ways, but with no explicit attempt to exploit synergies between activities.

Synergistic interventions include multiple components with the expectation that they will interact to produce an impact on the outcomes that is larger than the sum of the impacts produced by the single components independently implemented. For example, a poverty reduction programme may include the promotion of income-generating activities, but also health check-ups to reduce the number of out-of-work days to increase labour productivity. Similarly, a cash transfer programme may include nutrition education sessions in the hope that mothers will use the transfers to purchase nutritious food. This in turn will make children healthier, less likely to miss school, and more focused in the classroom.

Multicomponent programmes, with activities that reinforce each other, have a long history in development. They are rooted in the idea that development is multifaceted and that interventions complement each other. 'Integrated rural development', which was popular with international non-governmental organisations (NGOs) and aid agencies in the 1960s and the 1970s, is a typical example of this approach. In integrated development, communities or regions are simultaneously offered services in agriculture, education, health, and infrastructure. Recent examples of this type of intervention are the Millennium Villages Project (Masset, García-Hombrados and Acharya, 2020) and the Bangladesh Rural Advancement Committee (BRAC) Ultra Poor Graduation Programme (UPGP) (Bandiera *et al.*, 2017). Both interventions consisted of packages of health, education, and livelihood promotion. The core belief of these interventions was that improvements in different sectors would reinforce each other, thus moving households permanently out of poverty.

Critical-mix interventions are a simple variant of multicomponent interventions. In critical mixes, the intervention is successful only if the right combination of activities is implemented—a bit like in a recipe. The difference between this type of project and a synergistic intervention is subtle but relevant. In a critical-mix intervention, the different components have limited impact if implemented independently. The desired outcome is obtained if and only if all components are simultaneously implemented.

Box 1: The algebra of multicomponent interventions

Consider a project with two components x1 and x2 affecting the outcome y. A model to estimate project effects including those deriving from the interaction of the components takes the form:

$$y = a + bx_1 + cx_2 + dx_1x_2$$

The coefficient d measures the presence of a positive (or negative) interaction between components. In synergistic interventions, the d coefficient is different from zero and the interaction must be included in the specification.

In separable interventions the coefficient d equals to 0, there is no interaction between components and the effects of x1 and x2 can be independently estimated by the equation:

$$y = a + bx_1 + cx_2$$

In critical-mix interventions the coefficients b and c are both equal to zero. The components have no effect when independently implemented. They have an impact only when simultaneously implemented:

$$y = a + dx_1x_2$$

For an early example of this intervention, the 'Green Revolution' package consisted of the provision of improved seeds, fertiliser, and irrigation (Lipton, 2010). The large

outputs promised by the new high-yield varieties could not be obtained by using additional quantities of any of the three inputs independently. The three inputs would have to be adopted together and at the same time. Critical-mix packages are also common in nutrition and health.

3.2 Portfolio interventions

Portfolio interventions are projects bundled under a wider umbrella. They have become increasingly popular as a delivery method among donors and share a number of characteristics (Buffardi and Hearn, 2015). The projects in the portfolio are often implemented across sectors and across different populations. They are implemented at a large scale and over a long period of time. There is often an expectation that the projects will positively interact, although this is not a requirement. They have high-level goals that are broadly defined, often in relation to a specific thematic area such as empowerment, economic development, or climate change. While each project in the portfolio has its own theory of change, the overall portfolio has only a loose theory of change setting out an overall strategy to achieve the main goals. Since they involve many stakeholders and different intervention areas, they are very flexible and are adapted to specific contexts (Mackenzie and Hearn, 2016). Finally, they may require an overarching organisational structure that coordinates the various projects.

We further distinguish between three types of portfolio interventions: 'within-country portfolio interventions', which are implemented within a single country; 'cross-country portfolio interventions', which are implemented across multiple countries; and '*ex-post* portfolio interventions', which are considered as a portfolio after the fact.

Within-country portfolio interventions are similar to multicomponent interventions. They differ in that they are composed of different projects rather than of different component activities. At first glance, this can look like a difference in definition that has no real-world consequences, but the difference is significant. Components of the same intervention target the same agents in the same context, and are implemented by the same implementer. In a portfolio intervention, the projects will target different agents in different subcontexts and are implemented by different organisations. A portfolio intervention may even include an intervention that is itself a multicomponent intervention.

To clarify the distinction, consider the differences between a multicomponent intervention, such as BRAC's UPGP described above, and Foreign, Commonwealth & Development Office's (FCDO) Jordan Compact Economic Opportunities Programme (JCEOP). JCEOP is a portfolio of interventions designed '[t]o deliver the transformative

agenda set out in the Jordan Compact by: i) turning the refugee crisis into a development opportunity; ii) rebuilding Jordanian host communities; and iii) supporting Jordan's macroeconomic framework' (DFID, 2016). Whereas UPGP targets well-defined household-level outcomes, JCEOP targets system-level outcomes, some of which are left vague out of necessity. In order to cause changes in these high-level outcomes, JCEOP is designed as a package of concessional lending and technical assistance that enables and requires the Government of Jordan (through results-based payments) to carry out a suite of policy changes across many sectors. These include setting up special economic zones under a new investment law; policy changes to allow Syrian refugees to work and to own businesses; direct employment of refugees in public works schemes; and enrolling refugee children in regular Jordanian schools. The overarching theory of change shows the overall architecture of the portfolio, but does not describe the interventions that make up the portfolio. The precise details of the interventions are delegated by design to relevant branches of the Jordanian government.

Cross-country portfolio interventions are implemented in many countries at the same time, sometime in different continents. They differ from within-country portfolio interventions in some respects. First, they often need a super-national coordinating structure to monitor progress and direct activities. Second, they tend to have a narrower thematic focus than national-level interventions, being directed to address specific social problems such as malaria and undernutrition, as we discuss in the following examples. Third, the interventions implemented in different countries may also interact, although this is unlikely to be a goal, and some interactions will also take place at the level of coordinating the overall effort. Fourth and finally, they allow for a higher level of flexibility in implementation than national-level interventions because of the wider variety of contexts and opportunities for unforeseen policy changes.

For example, the US President's Malaria Initiative (PMI) is an interagency initiative of the US government led by the United States Agency for International Development (USAID) that implements a set of interventions for the prevention and case management of malaria (PMI, 2019). In each partner country, PMI works with the local government to develop and follow a Malaria Operational Plan that describes how the country will work to implement PMI's recommended interventions. Much like JCEOP, each country's PMI intervention will include sub-interventions across many sectors and at different system levels. PMI's recommended suite of interventions includes micro-level interventions like the supply of insecticide-treated nets, but also macro-level interventions like training, capacity strengthening of national malaria control programmes, and procurement systems reforms.

Feed the Future is another example of a cross-country portfolio intervention. Feed the Future is an initiative developed by the US Department of State and coordinated by USAID's Bureau for Food Security, but designed to be a 'whole of government' effort (Elliot and Dunning, 2016). Its mission is to 'combat global hunger, poverty and malnutrition'. In practice, this means USAID country offices or 'missions' in focus countries have developed multiyear strategies including the deployment of a variety of interventions to boost agricultural productivity and improve nutrition, particularly of women and children (Du *et al.*, 2015). These plans are guided by a shared approach but are heavily tailored to local conditions, in partnership with the local government and the private sector.

Ex-post portfolio interventions are bundles of interventions that were not designed to be part of an overall strategy, but that are considered as such by evaluators as a matter of convenience. For example, researchers and implementers may want to consider the overall impact of the 'portfolio' of interventions carried out by an aid agency, or the impact of the interventions carried out by a government agency in a specific country, or the impact of all development interventions by all agencies in another country. This definition of portfolio intervention assumes some level of coordination and strategy between the different projects considered, but this may be relatively low. Still, there may be interactions between the various interventions, even though they were not part of any specific programmatic strategy. The overall goals of the portfolio will be even less precisely defined than in the portfolio interventions described above.

An example of this approach is when the Independent Commission for Aid Impact (ICAI) was asked to review the relevance, effectiveness, and sustainability of UK overseas development assistance to Ghana relative to the UK's development objectives. Having decided upon a set of final outcomes of interest, ICAI assessed a large set of interventions for their contribution to those outcomes, despite the fact that the interventions had not been designed as part of a strategy targeting precisely those outcomes. In another example, ICAI carried out an assessment of all DFID's initiatives to improve maternal health. This 'thematic' portfolio bears a striking resemblance to the PMI portfolio previously discussed, although the interventions grouped under this portfolio were not created as part of a standalone initiative with a guiding strategy.

3.3 Interventions with long causal chains

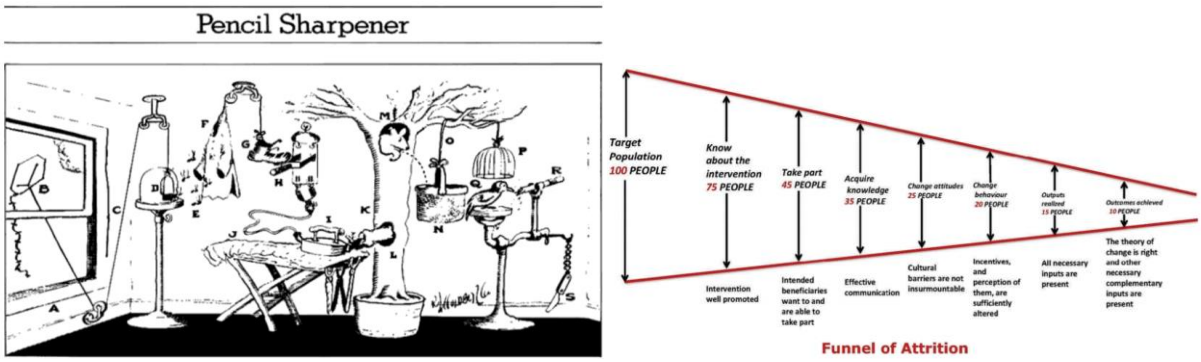
As discussed in the previous two sections, the complexity of multicomponent and portfolios interventions is driven by the number of activities, the number of target populations, and the number of project implementers. A different type of complexity arises from attempting to influence multiple behaviours. Some development

interventions are designed as long chains of causal steps, each of which requires influencing an agent to behave in a certain way. As in 'Rube Goldberg machines', these many steps in the theory of change present many opportunities to fail or for unexpected effects to emerge (Cartwright and Hardie, 2012). For example, in the Goldberg machine of Figure 2, a kite is operated by wind power to sharpen a pencil. The kite will sharpen the pencil only through a sequential and improbable chain of mechanisms. Some projects in international development have a similar structure in that they require a complicated sequence of steps to affect the final outcomes.

White (2018) describes a heuristic device he calls the 'funnel of attrition', which is useful for thinking through the effects of a long causal chain on final outcomes (right panel of Figure 2). In White's funnel of attrition, each step of the theory of change implies a reduction in the overall effectiveness of the intervention. In general, the project success will depend on the number of steps and on the difficulty of the task (behavioural change) produced at each step. Crucially, if one of the steps fails to operate, the whole machine breaks down.

Examples of this type of project abound in the development literature. Consider for example Educate! Experience, a leadership and social entrepreneurship skills training programme implemented by Educate!, an NGO working in Kenya, Uganda, and Rwanda. The programme teaches skills to students in their last two years of secondary school, and seeks to improve their level of skills, economic outcomes, educational attainment, and community involvement, as well as gender empowerment outcomes, social norms, and intimate partner violence outcomes (IPA, 2020). The first step in the theory of change underlying this intervention involves the NGO persuading a secondary school, or a local government controlling several schools, to partner with them. This eventually leads to skills training sessions being delivered in schools, which (all being well) leads to an increase in skills. There are then very many causal steps between increased skills and ultimate economic outcomes, community involvement, and intimate partner violence outcomes.

Figure 2: Development projects as Goldberg machines and the funnel of attrition



Sources: Cartwright *et al.* (2012) (left figure), and White (2018) (right figure)

In principle, each step of a project of this type should be tested before implementing the overall intervention. A sensible approach would consist of assessing the feasibility of each single link before starting overall activities. In practice, this is rarely carried out, and it is hardly feasible given the large number of sequential links. Instead, project developers rely on more or less explicit 'assumptions' about people's behaviour and response to interventions that may be described in diagrammatic 'theories of change'. Without this prior assessment of the effectiveness of each link in the chain, tracking the origin of the project's success—or lack of success—after the project has been completed becomes extremely hard. Process tracing has been proposed as a method to recover causal mechanisms of interventions in detailed case studies (Beach, 2017).[1]

An additional complication in the operation of projects with long causal chains is the presence at each step of the chain of 'supporting factors' and 'derailers', also defined as 'barriers' and 'enablers'. Supporting factors are features that are helpful or necessary for the programme to work as hoped, while derailers are features that might frustrate or stop a given causal step. Cartwright *et al.* (2020) discuss how each step in the causal chain requires a combination of factors in order to succeed, which is a stronger requirement than the linear process delineated in the funnel of attrition outlined above. For example, in the case of the Educate! Experience programme, some of the early causal steps in the theory of change will be enabled or frustrated by features of the school in which the intervention is implemented; students will only gain skills if the programme is implemented properly, the materials are made available, the school places a priority on the lessons, etc. Later steps in the theory of change may be frustrated or enabled by features of the economic and social context. The complexity of a causal step is therefore further enhanced by the number of additional features that either support or derail the activities. To sum up, the longer the causal chain and the

[1] We do not cover process tracing in our review, partly for reasons of space and partly because it falls outside the definition of causal inference method adopted.

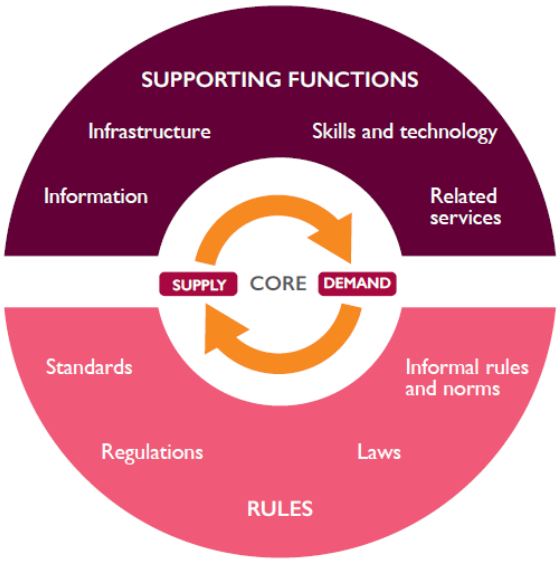
larger the number of supporting factors required at each step, the more complex the long causal chain intervention becomes.

3.4 System-level interventions

Recent years have seen a rise in popularity of 'system approaches' to project design. Packages of interventions have been implemented using a 'market system' approach, an 'education system approach', or a 'health system' approach. These interventions are not focused on service delivery or behavioural change, and set themselves the goal of changing the operation of the whole 'system'. They target different population groups at different levels in various ways. They often do not have a prespecified theory of change, as the characteristics of the interventions are developed during project operations while implementers learn how the project works.

A key element of these interventions is the development of a prior understanding of the system in which the project operates. For example, a popular system approach is Making Markets Work for the Poor (M4P), which has been supported by FCDO and the Swiss development agency (Springfield Center, 2015). The central idea of the M4P approach is that the poor depend on market systems for their livelihoods and that improving and changing market systems is key to poverty reduction. Markets are seen as embedded within a market 'system' including other markets and supporting factors such as education, health, water, and sanitation. The efficient and sustainable operation of markets in the system depends on the presence of basic services and the effectiveness of other interconnected markets. The conventional economics understanding of markets is represented by the familiar 'core' model of demand and supply in the centre of the diagram in Figure 3, where all market exchanges take place. However, it is argued, such exchanges cannot take place efficiently and sustainably in the absence of a set of rules and other supporting factors such as infrastructure, the flow of information, and the overall business environment. Once market systems are understood and all the interconnections with the surrounding environment are spelled out, specific interventions can be designed to remove the constraints that prevent the efficient operation of the system.

Figure 3: The M4P 'system' approach



Source: Springfield Centre (2015)

Projects following this approach are multifaceted, involve a different range of stakeholders, operate in several markets, and cover large geographic areas. For example, the FinMark Trust in South Africa aimed to provide financial services for the poor but, rather than offering loans to small farmers, it operated by building a shared vision of the industry, by supporting service innovation, by developing information services, by improving regulatory processes, and by stimulating consumer education. Similarly, the Katalyst programme in Bangladesh aimed to improve the productivity of vegetable farmers but, rather than provide technical assistance through conventional agricultural extension, it supported input suppliers to provide a retailer training programme.

System-level interventions are more clearly linked to complexity theory and complex phenomena. At the simplest level, system-level interventions change one system component at a time. These interventions, however, because they are based on an understanding of how the system works, will be aware of the prevailing trade-offs: how changes in one part of the system might deprioritise changes in another part. They will also consider spillover effects and displacements: how the solution of a problem in one area creates problems in other areas.

More often, system-level interventions want to change multiple elements of a system at a time. This arises from an understanding of the system as a complex system dominated by feedback loops and emergent phenomena. In the presence of loops, an intervention may change the system in such a way as to affect the effectiveness of the intervention by accelerating its impact or by dampening it. In the presence of emergent

phenomena, a change in the system may completely change the way the system operates. These types of intervention are therefore sometimes directed to produce a seismic and qualitative change in the way the whole system operates.

Table 1 present a summary of the complexity of each intervention type along a series of characteristics. These include:

1. the number of project components: interventions become more complex as the number of component activities increase;
2. the number and difficulty of behaviours targeted along the causal chain: the longer the causal chain and the more sensitive it is to contextual characteristics, the less likely the project is to achieve its goals;
3. the number of groups and organisation targeted: interventions become more complicated as they target different population groups in different ways;
4. the number of sectors: simple interventions address a specific social problem, and complex interventions may address wide social problems encompassing several sector areas at the same time;
5. the number of stakeholders: as interventions target different population groups and across different sectors, they involve a plurality of stakeholders and funders. The coordination of sometime conflicting ambitions of different stakeholders adds an additional layer of complexity;
6. the scale of the intervention: interventions are more complex when they are implemented over a large geographic area, such as a region, a country, or a number of countries;
7. the level of standardisation: complex interventions tend not to state precise or quantifiable goals. The activities are untested or being piloted and the outputs are uncertain; and
8. the presence of emergent phenomena: a higher level of complexity is reached when interventions include synergistic interacting components and when their operation is driven by feedback loops and non-linear changes.

Table 1: Main characteristics of complex interventions

	Multicomponent interventions	Portfolio interventions	Long causal chain interventions	System-level interventions
Number of components				
Number and difficulty of behaviours				
No. of groups and orgs. targeted				
Number of sectors				
Number of stakeholders				
Scale of intervention				
Level of standardisation				
Emergent outcomes				

Note: Red cells represent a strong association between intervention type and complexity feature. Orange squares represent a moderate association. Green squares represent low association.

The characteristics in Table 1 are colour-coded for increasing level of complexity, where red represents the highest level. The coding is designed to help draw some conclusions on the four project types discussed above; it is not intended as a tool for managers and researchers to assess the degree of complexity of an intervention. Bamberger *et al.* (2016) built a checklist covering the following four dimension of complexity: the nature of the intervention, the interactions between institutions and stakeholders, causality and embeddedness, and the nature of the system. The checklist can be used to rate the complexity of any intervention on a scale. We agree that the complexity of an intervention is a matter of degree, and that a checklist is a useful tool to guide decisions of managers and researchers. Our interest, however, is simply to identify the main patterns and differences between project types.

A glance at Table 1 suggests that interventions with long causal chains are the least complex among the four types considered. Interventions with long causal chain intervene in several complex social processes and require a large number of supporting factors, but they are highly standardised. They do not necessarily include multiple components and are normally implemented in a single sector and a circumscribed geographic area, thus not involving multiple stakeholders. Finally, they are not expected to produce emergent outcomes by design.

Multicomponent interventions are more complex than long causal chain interventions because carrying out many activities at the same time implies the involvement of many stakeholders, various sectors of intervention, and multiple targeted groups.

Multicomponent interventions are often designed to produce emergent outcomes, such as synergies between components. However, they are normally implemented at a local level and, having a predefined set of goals and means to achieve them, they can be standardised.

Portfolio interventions and system-level interventions present the highest level of complexity, including the possibility of generating emergent outcomes. An important distinction also emerges here between localised projects, which are implemented in a geographically limited area, and interventions that are implemented at a national level or cross-country. Multicomponent interventions and long causal chain interventions normally belong to the first category, while portfolio interventions and system-level interventions belong to the second. The scale of the intervention has a large impact on the difficulty of evaluating an intervention and on the number of methodological options available, often superseding other considerations. Portfolio and system-level interventions are, for this reason alone, far more complex and difficult to evaluate than other projects.

Based on this brief description, it should come as no surprise that portfolio interventions and system interventions are the most difficult to evaluate. Evaluations of long causal chain interventions are fairly common, and several methods are available to deal with the number and the difficulty of the behavioural changes they produce. Similarly, multicomponent interventions are often evaluated, albeit in some special cases and with a limited number of methods. However, portfolio and system-level interventions are rarely evaluated, despite being very common in the practice of development, and methods for their evaluation are lacking. The evaluation challenges posed by each project type and current approaches to their evaluation will be the subject of the next section.

4. Challenges of evaluating complex interventions and current approaches

In this section, we summarise the main challenges encountered in the evaluation of complex interventions and the prevailing evaluation approaches adopted for each intervention type. In the previous section, we defined the complexity of interventions as characterised by the number of activities, target groups, sectors of interventions, and stakeholders; the difficulty of targeted behaviours; the geographic scale; the level of standardisation; and the presence of emergent outcomes. Each of these characteristics poses big challenges to evaluators. Some of these challenges can be classified under the traditional headings of internal and external validity, while others are of a more practical nature and relate to measurement and logistics. There are also issues specific to complex interventions alone, which relate to the difficulty of evaluating impacts on emergent outcomes and adaptive interventions.

The multiplicity of components, target groups, sectors, and stakeholders complicates the task of understanding the impact of single activities and ultimately of understanding how the intervention operates. Evaluators normally obviate this by using either of the following strategies. The first consists of singling out components in a specific sector and evaluating that component, thus ignoring how the overall intervention operates. The second strategy consists of evaluating the whole intervention as a 'package', without investigating the operating mechanisms. Alternatives to these two strategies include building (or analysing) multiple treatment groups in experimental designs, an approach that is practically difficult to implement and very expensive. The multiplicity of inputs normally also translates into a multiplicity of outcomes, which is the source of additional problems regarding the statistical significance and the correct reporting of results.

The presence of multiple stakeholders adds a political dimension to complex interventions. Many complex interventions are the result of compromises between different actors, governments, and institutions. These various agents compete and negotiate for the use and distribution of the resources involved. This translates into the difficulty of defining the outcomes of the intervention. Different agents will have different goals and will value the same outcomes in different ways, in such a way that an objective and shared evaluation of the intervention among the various agents is difficult to obtain.

The difficulty of targeted behaviours in long causal chain interventions complicates the task of explaining the cause of success and of failure. This problem is often addressed by conducting a mixed-method evaluation, giving a significant role to qualitative analysis of implementation and to quantitative monitoring data, for example through process evaluations. The complexity of long causal chain interventions is also the source of a large sensitivity of results to contextual factors. Impacts of these interventions will vary greatly with the presence of different supporting factors and derailers in such a way that impacts cannot be easily extrapolated from one context to another and are hard to predict.

The large scale of interventions is an obvious problem for evaluation approaches that rely on counterfactual analysis. Valid control groups are hard to find when interventions are implemented at the national level or across multiple countries. Large-scale interventions are also often implemented over several years and may produce outcomes only after a long period of time, but standard evaluations are designed to observe project effects after two or five years. In addition, long-term effects—for example on empowerment or employment—might be difficult to measure or to disentangle from other effects.

Complex interventions, particularly those operating at the system level, are not standardised and tend to be adaptive. These interventions do not follow a precise protocol when implemented. Theories of change and logical frameworks of adaptive interventions tend to be loose (Davies, 2018). Intervention designs indicate strategies rather than protocols of action. There are therefore challenges related to what outcomes and inputs should be considered, and how these can be defined before the intervention in prospective evaluations. Adaptive interventions may have impact in unexpected areas and can produce spillover effects that are difficult to anticipate. In addition, as these interventions modify or drop activities altogether in the course of implementation, they require the collection of timely data that monitor actual implementation and outcomes using special monitoring data. Programme managements are often unable to adjust to emergent and unexpected events. Many management systems do not have the ability to absorb real-time data and make the necessary adjustments. Programmes also vary in their ability to adjust. For example, infrastructure interventions are by design less likely to be able to modify project design during implementation.

Table 2: Main challenges in evaluating complex interventions

Complexity	Challenge
Multiplicity of inputs, target groups, sectors, and stakeholders	• Assessing impact of single components • Reporting and interpreting multiple outcomes
Difficulty of targeted behaviours	• Understanding impact of intervention • High contextual sensitivity of results
Geographic scale	• Absence of valid control group • Long-term and unmeasurable outcomes
Lack of standardisation	• Prospectively design an evaluation • Availability of timely data
Emergent outcomes	• Estimation of non-linear outcomes, synergies, and multiple equilibria

A final challenge refers to the generation of emergent outcomes. These include complex non-linear responses. For example, impacts may not be visible for a long time and may suddenly appear exponentially. Other complex interventions may aim at generating systemic changes, for example those predicted by poverty trap theories or by other theories of multiple equilibria. Even simpler multicomponent interventions can produce emergent effects in the form of input and output complementarities. Emergent outcomes, such as multiple equilibria, synergistic effects, and exponential changes, are difficult to observe and measure. Standard evaluation approaches are based on the assumption that such phenomena do not arise. For example, RCTs assume the stable unit of treatment value (SUTVA): this means that impacts on an agent are independent of impacts on other agents, which rules out the emergence of outcomes through agents' interactions. Table 2 summarises the main challenges in the evaluation of complex interventions described so far.

4.1 Approaches to evaluating multiple component interventions

In order to obtain examples of current approaches to the evaluation of multicomponent interventions, we accessed the 3ie repository of impact evaluations. As of December 2020, the 3ie repository included 3,754 impact evaluations of interventions in low- and middle-income countries published since 1990. The repository only includes studies that employ an experimental or quasi-experimental design as a main evaluation approach or in combination with other methods. The repository includes non-counterfactual methods only if employed in conjunction with a counterfactual approach, which limits the range of methods covered. We searched studies that included, either in the title or in the abstract, at least one of the following words: 'complex', 'multicomponent', or 'package'. The search returned 84 valid studies from the

health sector (61%); the social protection sector (12%); the education sector (8%); and the agriculture sector (7%). The remaining studies were from the financial sector; the water, sanitation, and health sector; and the energy extractives sector.

The most common evaluation approach among the selected studies (61%) consisted of assessing the multicomponent intervention as a 'package'. Researchers adopting this approach do not attempt to estimate the impact of single components of the project. Rather, they assess the impact of the whole package at the end of the programme. This approach is useful for accountability purposes, but it is of limited use for policy. Since the mechanisms of the interventions remain unexplored, the approach does not allow for the identification of the most effective component or of the most effective combination of components. For the same reason, the results of these evaluation cannot be easily extrapolated to other contexts and have limited external validity. A partial remedy to this problem is to run multisite trials whereby the same intervention is implemented in many contexts at the same time (Banerjee *et al.*, 2015), summarising the results through meta-analyses or by exploring the heterogeneity of the results. But multisite trials are expensive and the standardisation of the intervention across contexts, which is required for the extrapolation of the results, is hard to achieve in practice. A similar approach has also been promoted by the Evidence in Governance and Policy group through the *Metaketa* initiative. The aim of *Metaketa* is to address a question of policy importance for governance by coordinating, in a collaborative way, a cluster of field experiments that are independently carried out.[2]

The second most common approach (30%) consisted of setting up a multi-arm randomised design, in which different arms received different interventions. In most cases, multi-arm studies analyse the impact of different interventions separately, but they allow for testing the effects interactions. The majority of multi-arm trials in the 3ie database consisted of 'cross-cutting' designs (Duflo, Glennerster and Kremer, 2007). In a typical cross-cutting design, half of the sample receives the treatment and the other half is the control group. The researchers then further randomise a variant of the intervention within the treatment group so there are three arms: two treatment groups and a control group. These designs allow testing different interventions and their interactions, but this is rarely done in practice and researchers often report results for only one, or a few, components and ignore the interaction effects.

Muralidharan *et al.* (2019), found that, out of 27 cross-cutting design evaluations in the top five economics journals between 2007 and 2017, only eight estimated all

[2] Lessons learned from the *Metaketa* initiative can be found in the book edited by Dunning *et al*. (2019).

component interventions and their interactions. Given that the full sample of RCTs considered consisted of 124 studies, this is a relatively small fraction of the full sample of studies. In most cases, this occurred because the sample size of the studies did not allow for testing the effects of interactions, highlighting the fundamental limitation of multi-arm studies in analysing multicomponent interventions: they need extremely large samples, often beyond the budget availability of most researchers.

A small number of studies in the 3ie database (6%) tried to identify the effects of the different project components using quasi-experimental methods. A study of HIV prevention through different types of condom promotion used matching methods to analyse the impact of different components and of their interactions (Juneja *et al.*, 2013). A study of a multifaceted nutrition intervention in Niger employed instrumental variable methods to unpack the impact of single components (Brück *et al.*, 2019). Another study proposed using regression analysis to assess the impact of different components with an example application from education interventions in Zambia (Elbers, Gunning and de Hoop, 2009; Elbers and Gunning, 2014). The quasi-experimental approaches employed in these studies do not have any advantage over RCTs in terms of the estimation of different components and of their interaction, although they can be carried out using a smaller sample size and at a lower cost. Their results are less credible than those produced by RCTs because of the biases inherent in quasi-experimental methods. We are not aware of a quasi-experimental method that is able to assess the effects of different components of an intervention, including their interactions, in an unbiased way.

Finally, a few studies (3%) explicitly considered only one component of the intervention in isolation from the others. We suspect this latter approach is much more common than it appears from our quick search. Development projects often consist of several components. However, researchers often focus on a single component of the intervention, thus implicitly assuming that the effects of this component are independent of the effects of the other components.

Beyond breaking programmes into single components and their interactions, the studies identified by our search addressed the complexity of the interventions by conducting parallel qualitative studies detailing the implementation of the interventions. Typically, these studies were conducted alongside process evaluations that investigated the fidelity of the intervention to the project design and how the different components were implemented. These qualitative studies produced suggestive evidence on the relative effectiveness of different components and on their interactions, but did not inform the design of the quantitative evaluation. For this

purpose, they should rather be conducted as formative assessments at the design stage of the studies.

4.2 Approaches to evaluating portfolio interventions

The 3ie repository of impact evaluations does not contain studies evaluating portfolio interventions. The 3ie database only includes studies using either experimental or quasi-experimental designs alongside other methods, which are not suitable for the evaluation of portfolio interventions. To identify evaluations of portfolio interventions, we adopted a different approach. We first listed major bilateral and multilateral organisations and philanthropic foundations.[3] We then searched the websites of each organisation to identify large projects that corresponded to our definition of portfolio intervention. Finally, we searched the literature to identify evaluations of the same interventions.

Table 3: Evaluations of portfolio interventions

Project name	Funder	Authors	Method
Determined, Resilient, Empowered, Aids-free, Mentored and Safe partnership to reduce HIV among adolescents and young women	USAID	Birdthistle *et al.* (2018)	Before–after analysis
Feed the Future	USAID	Ryckman *et al.* (2020)	Difference-in-difference
Malaria Vaccine Development Program	USAID	Nsubuga *et al.* (2016)	Narrative synthesis of desk reviews and key informant interviews
The President's Emergency Plan for HIV Relief	USAID	IOM(2013)	Contribution analysis, and narrative synthesis of desk reviews and key informant interviews
PMI	USAID	USAID (2011)	Narrative synthesis of desk review and key informant interviews

[3] The organisations considered were: EU, FCDO, GIZ, JICA, USAID, AfDB, FAO, IADB, IBRD, IFAD, ILO, UNICEF, WFP, WHO, the Ford foundation, the Gates foundation, the Hewlett foundation, the Macarthur foundation, and the Rockefeller foundation.

Project name	Funder	Authors	Method
Poverty Reduction Support Facility	Australian Aid	Mackenzie and Hearn (2016)	Non-counterfactual methods (contribution analysis, general elimination methodology, etc.)
Global Program to Acceleration to End Child Marriage	UNICEF	UNFPA–UNICEF (2019)	Narrative synthesis of desk reviews, key informant interviews, and secondary data analysis
Transforming Health Systems Initiative	Rockefeller Foundation	Smith *et al.* (2018)	Narrative synthesis of key informant interviews and of secondary data analysis
Big Bet on Nigeria	MacArthur Foundation	Franco *et al.* (2019)	Before–after analysis
Affordable Medicine Facility for Malaria	Global Fund	Tougher *et al.* (2014)	Before–after analysis

In this way, we identified 10 evaluations of portfolio interventions (Table 3). This is a limited sample of the types of evaluation considered, but we believe it is illustrative of prevailing approaches. Portfolio projects are ambitious and large, and they are normally evaluated, if only for accountability purposes. However, evaluations are rarely published in academic journals and often consist of reports.

Evaluations of portfolio interventions normally include a plurality of methods. It seems sensible that projects operating at different levels, with impacts on different sectors and countries, should be assessed in ways that are the most appropriate in each case. For example, researchers may want to know the level of coordination between stakeholders, the overall impact of an intervention on welfare outcomes, the political process that influenced implementation, the impact of the project on public finances, and so forth. All these evaluation questions require different methods.

In our review, the majority of evaluations employed a combination of desk reviews, stakeholder and partner surveys, key informant interviews, and field visits. Qualitative

observations were then used to build a qualitative narrative of impact that led to a number of lessons learned. The narrative was often supported by the analysis of secondary data. These could be project monitoring data or socioeconomic trends. These qualitative analyses were conducted at what appears to be a relatively low cost. It is indeed a bit surprising that funders devote limited resources to the evaluation of interventions that are implemented at a very large scale and that involve huge disbursements.

Some studies employed qualitative methods of analysis that are not based on counterfactuals. These include, for example, process tracing, general elimination methodology, and developmental evaluation. A popular choice was contribution analysis (Mayne, 2001). Contribution analysis is based on two key assumptions. The first is that a project is only one of several factors contributing to determine observed outcomes. The second is that it is impossible to attribute observed impact to a single intervention. The method explores plausible associations between observed impacts and the programme by building theories of change and by collecting evidence to corroborate different causal pathways of impact. The method is useful for building a theory of the intervention and for identifying concurring causes, supporting factors, and alternative plausible causal pathways, but it does not alone assess the effectiveness of the intervention (White, 2010).

Quantitative approaches to evaluating portfolio interventions focus on the long-term and on high-level outcomes at the country or cross-country level. For example, some studies in our review assessed the impact of interventions on prices, HIV, and malaria incidence. Before–after comparisons are common. In a before–after analysis, outcomes are measured before and after the intervention. Researchers are aware that a comparison of outcomes before and after a programme is likely to be biased by other factors occurring at the same time, but the analysis is nevertheless conducted as a first approximation of the estimation of project effects.

In some cases, a difference-in-difference analysis is conducted, which produces more plausible results. In this type of study, the interventions are analysed as natural experiments using the variation in implementation and outcomes across countries. Difference-in-difference analyses provide relatively weak evidence of impact and are sometimes difficult to carry out because of the difficulty of identifying a valid control group. It is a bit surprising, however, that this method is not more often employed. Many United Nations agencies, for example, implement relatively standardised projects in a plurality of countries while collecting data on outputs, outcomes, and project

performance ratings. Datasets are available for evaluating a number of these interventions at a relatively low cost.

4.3 Approaches to evaluating long causal chain interventions

Interventions with long causal chains are very common. Impact evaluations of long causal chain interventions are also extremely frequent. Few impact evaluations today consist of simple measurements of outcomes in a project and a control group after an intervention. Researchers are required to explore and investigate the mechanisms behind the operation of interventions. This happens in different ways. Typically, randomised experiments are conducted alongside rigorous process evaluations that analyse the fidelity of the intervention to the project design and help map out the outcomes of the interventions along the project causal chain. Other theory-based approaches, and structural modelling in particular, are designed to disentangle and separately analyse the various links between the factors determining the operation of a project. Other evaluations consist of 'mechanism experiments', whereby the goal of the researchers is not directly to assess the impact of an intervention but to understand the operating mechanism behind one or more links of the project causal chain. More simply, researchers often assess the impact of interventions over many outcomes, with the goal of mapping the effectiveness along the causal chain.

Methods for evaluating long causal chain interventions are available. In some cases, the intervention can be unpacked in the form of a result chain, and the project effects (if data are available) can be estimated at each step of the chain with experimental or quasi-experimental methods (see White and Masset (2007) for an example application of this approach to a complex nutrition project in Bangladesh). Mechanism experiments have been proposed to test the effectiveness of single steps or of theoretical hypotheses supporting a single step of the chain. This approach appears to be particularly valuable when there is uncertainty regarding the effectiveness of a specific step within an overall result chain, whose remaining pieces of the mechanism are known to work (Ludwig, Kling and Mullainathan, 2011). As mentioned, however, the most popular approach consists of running 'realist' or 'pragmatic' RCTs, in which the evaluators formulate testable hypotheses about the operation of the project mechanism and of its interaction with the context and then use a combination of outcome and process data to analyse meditation and moderation effects (Jamal *et al.*, 2015).

Each of the above approaches will perform better or worse than others in specific circumstances. Overall, we believe the methodological problem of evaluating interventions with long causal chains is fundamentally solved. Qualitative and

quantitative methods that are fit for purpose are available, and researchers are trained in understanding and testing the mechanisms operating behind this type of interventions. The same cannot be said about the external validity of these evaluations. Although the results of realist evaluations, mechanism experiments, and theory-based evaluations can be used to infer the effectiveness of interventions in different contexts, none of them can claim the ability to extrapolate results to other contexts, which remains a lively area of methodological research.

4.4 Approaches to evaluating system-level interventions

One typical approach to understanding the functioning of complex systems is the computable general equilibrium (CGE) approach used in economics. In this approach, the economy is modelled as a complex system of interacting parts and agents. The interrelations between the components are specified by equations. The system is shocked by a hypothetical intervention, and the impact on the variables is observed once a new equilibrium is obtained. The method was originally developed to model national or world economies, but it can equally be used to model small economies at the regional or village level (Taylor and Filipski, 2014). A search of the term 'general equilibrium' in the 3ie database returned 187 hits. Many of the results were standard evaluations discussing general equilibrium effects, but others explicitly employed micro-level CGE to evaluate the impact of interventions. For example, CGEs have been used to evaluate the spillover effects of cash transfer interventions in village economies (Thome *et al.*, 2013) and the impact of cash transfers to refugees on the economy of the host countries (Taylor *et al.*, 2016).

A limitation of CGEs is that they are equilibrium models and not designed to model the complex emergent outcomes that system interventions set out to affect. In order to access the literature on the evaluation of complex system interventions, we search two publications—the *American Journal of Evaluation* and *Evaluation*—over the period from 2000 to 2020 using the word 'system' either in the title or in the abstract. Some of the studies identified were entirely conceptual. They discussed the difficulty of evaluating interventions implemented in complex system, particularly of experiments and quasi-experiments, and suggested methods based on system analysis such as 'system dynamics' (Fredericks, Deegan and Carman, 2008), 'critical systems heuristics' (Gates, 2018), 'complex systems lens' (Moore, Parsons and Jessup, 2019), 'epistemological pluralism' (Jones *et al.*, 2016), 'system thinking' (Hummelbrunner, 2011), 'higher-level theory' (Barnes, Matka and Sullivan, 2003), 'complexity-consistent theory' (Barnes, Matka and Sullivan, 2003), and 'critical system thinking' (Reynolds, 2014).

We also found a number of applied studies employing system approaches to support the evaluation of specific interventions. Several of these projects proposed the use of specific methods, such as 'complex system informed theories of change' (Zazueta, Le and Bahramalian, 2021), 'outcome evidencing' (Paz-Ybarnegaray and Douthwaite, 2017), and 'actor-based change framework' (Koleros *et al.*, 2020), with the goal of building theories of change that are complexity-aware. Other studies proposed methods, such as 'reflexive process monitoring', for monitoring complex interventions (van Mierlo, Arkesteijn and Leeuwis, 2010), or theory-based approaches more broadly to assist the design of complex interventions (Koleros *et al.*, 2016).

None of these studies evaluated the effectiveness of the interventions considered. Rather, they proposed and employed methods to better understand the complex system in which the projects operated and to build better conceptual frameworks and theories of change. The goal of these studies was not to assess the effectiveness of interventions, but to inform research and data collection, to formulate research hypotheses, and to interpret empirical results. As suggested by Moore *et al.* (2019), the system approaches described in these studies are better considered as aimed at framing relevant questions and interpreting results. They should preferably be used at the design stage of an intervention, or they should be employed to support evaluations conducted using other methods rather than as standalone evaluation approaches.

McGill *et al.* (2021) reviewed the literature on system methods in public health, defined as approaches adopting a complex systems perspective to evaluation, and described and classified the following methods: system mapping, network analysis, system modelling, and system framing. System mapping methods, such as 'causal loop diagrams' and 'concept maps', are used to illustrate the structure of a system and how its different components interact. They are often based on stakeholders' consultations and other qualitative data. They are valuable to understand the key points of intervention in a system, and therefore help project design. Network analysis methods, such as 'sociograms' and 'viable systems models', focus on the relationships between individuals in a system. They are used to examine issues emerging in qualitative research and to investigate phenomena such as the diffusion of innovations. System modelling approaches, primarily 'agent-based modelling' and 'system dynamics', are computational methods that simulate the behaviour of a system over time. They are used to make predictions about the outcomes of hypothetical interventions or about the long-term outcomes of real interventions under different scenarios. System framing involves methods that try to apply system theory and concepts to other evaluation methods. These include QCA, process evaluations, and concept diagrams. They are

often carried out through stakeholder consultations to understand how the components of an intervention interact with other elements of the system.

These approaches are useful at theorising, predicting, and framing issues from a complexity perspective. They are useful when the goal is to conceptualise the interactions between the various components of a system. They allow formulating a hypothesis about the evolution of a system and performing predictions about its behaviour under different scenarios. They are therefore extremely useful at the design stage of an intervention, but they are not designed to assess the effectiveness of interventions in the form of impact evaluation studies. Some of these methods can be potentially used in impact evaluations, but they will have to be refined or associated with other methods (as we will discuss in Sections 5.6 and 5.7).

Table 4 summarises the approaches to evaluating complex interventions that were discussed in this section.

Table 4: Common approaches to evaluating complex interventions

Complex intervention	Popular approaches
Multicomponent interventions	• Multi-arm trials • 'Package' evaluations
Portfolio interventions	• Desk reviews, key informant interviews • Contribution analysis • Before–after and difference-in-difference
Long causal chain interventions	• 'Realist' and 'pragmatic' RCTs • Mechanism experiments
System-level intervention	• General equilibrium analyses • Qualitative 'system' approaches

5. Underutilised approaches for the evaluation of complex interventions

In this section, we review a series underutilised evaluation methods that we believe are suited to evaluating complex interventions and that address causality. These methods are factorial designs, adaptive trials, QCA, synthetic control, agent-based modelling, and system dynamics. They were identified through reading the literature on the evaluation of complex interventions discussed in the previous section.

5.1 Search strategy and results

We set out to review the use of factorial designs, adaptive trials, QCA, synthetic control, agent-based modelling, and system dynamics in the evaluation of development interventions. To do so, we searched the following sources:

- method systematic reviews;
- the 3ie repository of impact evaluations;
- specialised academic journals; and
- Google searches.

We started with searching for the selected methods within existing systematic reviews. We searched for systematic reviews using the Google search engine by screening the first 10 pages returned after entering key terms. For example, for adaptive trials, we used the search terms '*Adaptive trial systematic review*'; '*Adaptive trial systematic review agriculture*'; '*Adaptive trial systematic review education*'; '*Adaptive trial systematic review health*'; and '*Adaptive trial systematic review environment*'. A similar approach was followed for the other methods. We used the reviews to identify primary studies that had been identified by previous searches.

We obtained primary studies by searching the 3ie repository of impact evaluations (https://developmentevidence.3ieimpact.org/). The 3ie repository includes impact evaluations of development interventions that employ an experimental or quasi-experimental design as their main evaluation approach or in combination with other methods. As of January 2021, it included 3,754 impact evaluations of interventions in low- and middle-income countries published since 1990. We used the repository search engine using key words such as 'factorial' and 'adaptive'.

Other primary studies were found by consulting specialised journals. In particular, we searched the Comparative Methods for Systematic Cross-Case Analysis (COMPASSS)

repository for QCA studies. COMPASSS (https://compasss.org/about/) is a network of scholars and practitioners who share a common interest in QCA. It was launched in 2003 and maintains a bibliography (https://compasss.org/bibliography/) of all journal articles on QCA, which is regularly updated. We consulted all articles published by the *Journal of Artificial Societies and Social Simulation* (JASSS) to obtain agent-based modelling studies. JASSS is an interdisciplinary journal for the exploration and understanding of social processes by means of computer simulation. It was founded in 1998 and includes the largest selection of agent-based modelling studies available. Finally, we consulted the Google search engine for additional searches. In particular, we used Google to identify systematic reviews, but also unpublished papers and reports belonging to grey literature.

We screened primary studies for eligibility using two simple inclusion criteria:

- the study ought to be conducted in a low- or middle-income country, as defined by the World Bank classification; and
- the study should evaluate a development intervention (thus excluding conceptual papers, reviews, and the like).

Screening was conducted independently by two reviewers in two stages, and contradictions were resolved through further analysis of the papers. Table 5 shows the number of papers for which the title and abstract were screened (second column) and the papers eventually selected for the review for each method (third column). The numbers in the second column refer to the studies identified by the search based on title and abstract, and by applying the two eligibility criteria above. The third column shows the number of studies ultimately selected after removing studies that, on closer examination through full-text screening, did not meet the eligibility criteria. Further details of the screening process can be found in Appendix 1.

Table 5: Primary studies identified by the review

Method	Title and abstract review	Final selection
Factorial design	81	27
Adaptive trials	9	2
QCA	124	8
Synthetic control	23	8
Agent-based modelling	76	11
System dynamics	34	8

5.2 Factorial designs

Factorial designs are randomised experiments that assess the impact of different treatments and of their interactions. They are sometimes confused with multi-arm trials, but their goals, characteristics, and interpretation are different. In this section, we first describe factorial designs and then we discuss our review of the use of factorial designs in the evaluation of development interventions. Finally, we conclude with some comments on the range of applicability of this approach.

Factorial designs were originally developed in agriculture. They have been used in engineering, psychology, and marketing, but rarely in the social sciences. They are often confused with multi-arm randomised trials, despite being very different (with the exception of some rather special cases). In a multi-arm trial, researchers are testing more than one intervention at a time. For example, they may want to assess the effectiveness of intervention A and of intervention B against a control group representing the status quo. To carry out the experiment, the researchers define three groups: a group receiving intervention A, a group receiving intervention B, and a control group not receiving any intervention. Occasionally, researchers include a fourth intervention group comprising subjects receiving both interventions A and B in order to assess the combined impact of the two interventions. This latter version of a multi-arm trial is similar to a 2-by-2 factorial, which is the simplest of factorial designs. Factorial designs, however, come in many other forms that bear no similarity to multi-arm trials.

An exhaustive description of factorial designs can be found in Box and Hunter (2005), whereas Collins *et al.* (2014) and Baker *et al.* (2017) provide more accessible introductions. We start with a bit of terminology. In factorial designs, interventions are called 'factors' and are administered at different 'levels'. Most studies employ only two levels—for example, the intervention is implemented or the intervention is not implemented. Some factorial designs, however, use more than two levels, for example three levels—low, medium, and high—to assess responses to different dosages of the same treatment. The simplest factorial design is the 2-by-2 factorial, which has two factors and two levels. For example, the study by Mbiti *et al.* (2019) tests the impact of providing grants to schools (intervention A), the impact of teacher incentives (intervention B), and the impact of providing incentives and grants at the same time. The analysis of interactions between variables is what makes this type of study suitable for evaluating interventions that rely on complex relationships between activities, such as synergistic effects.

It is customary to design a factorial experiment using a table in 'standard form'. In this table, receiving the intervention is represented by a plus sign and not receiving the

intervention by a minus. For example, Table 6 is a table in standard form of a 2-by-2 factorial design. Usually, the signs of the first intervention considered are alternated the fastest, while the signs of the last intervention are alternated the slowest. It is also customary to start the table at the low levels of the factors and to finish at high levels for all factors. This is of limited significance in a 2-by-2 factorial, but the standard form becomes more useful in more complex designs.

Table 6: A 2-by-2 factorial in standard form

School	Intervention A	Intervention B	Y (test scores)
1	–	–	100
2	+	–	120
3	–	+	105
4	+	+	130

Table 7 is a 2-by-3 factorial experiment in standard form, with two levels and three factors. More complicated designs are possible by adding more factors. If L is the number of levels and if k is the number of factors, then the number of experimental units needed is L^k. For example, with four factors and two levels, the minimum number of experimental units is $2^4 = 32$. Note that the number of experimental units increases exponentially with the number of factors.

Table 7: A 2-by-3 factorial in standard form

School	Intervention A	Intervention B	Intervention C	Y test scores
1	–	–	–	100
2	+	–	–	120
3	–	+	–	105
4	+	+	–	130
5	–	–	+	110
6	+	–	+	125
7	–	+	+	110
8	+	+	+	130

A key difference between factorial designs and standard RCTs is that a factorial design does not need a large number of observations for each factor or combination of factors. For example, in the 2-by-3 factorial study of Table 7, 50% of observations receive intervention A, 50% of observations receive intervention B, and 50% receive intervention C. The whole sample can be split in two for estimating the impact of A, the impact of B, and the impact of C respectively. This is possible because the interventions overlap. In other words, a factorial design employs samples more efficiently and estimates impact of interventions using a smaller number of observations in comparison to a standard randomised trial. However, because of the overlap between interventions, the

interpretation of the estimated effects of a factorial design is different from the interpretation of effects of a standard randomised trial.

Box 2: Numerical example of main and interaction effects of factorial designs

Table 6 and of Table 7 shows the impact of the interventions on students' test scores. In the case of Table 6, we proceed to calculate the main effects and the interaction effects in the following way:

- The main effect of A is the average of the effect of A, when B is on and when B is off. This is: [(130-105)+(120-100)]/2=45/2=22.5.
- The main effect of B is the average of the effect of B, when A is on and when A is off. This is: [(130-120)+(105-100)]/2=25/2=12.5.
- The interaction effect is one-half the difference between the effect of A when B is on and off. This is: [(130-105)-(120-100)]/2=5/2=2.5 or, equivalently, it is half the difference between the effect of B when A is on and off—that is: [(130-120)-(105-100)]/2=5/2=2.5.

In the case of Table 7, main and interaction effects are calculated in the following way:

- Main effect of A=[(120-100)+(130-105)+(125-110)+(130-110)]/4=20.
- Main effect of B=[(105-100)+(130-120)+(110-110)+(130-125)]/4=5.
- Main effect of C=[(110-100)+(125-120)+(110-105)+(130-130)]/4=5.

The 2-way interactions are:

- AB interaction={[(130-110)-(125-110)]/2 + [(130-105)-(120-100)]/2}/2=2.5.
- AC interaction={[(130-110)-(130-105)]/2 + [(125-110)-(120-100)]/2}/2=-2.5.
- BC interaction={[(130-125)-(110-110)]/2 + [(130-120)-(105-100)]/2}/2=2.5.
- The three-way interaction is half the difference of the effect of AB at C high and at C low: ABC interaction = {[(130-110)-(125-110)]/2 - [(130-105)-(120-100)]/2}/2=0.

With more factors, the calculations become more and more complicated but are normally conducted using computer software available in packages such as Stata and R. The same results can also be obtained using OLS, but the variables need to be transformed before running a regression. With dummy variables, as in this case, the transformation consists of simply transforming the A+ and B+ variables into +1 and the A- and B- variables into -1. The transformation needs to be taken into account when interpreting the effect sizes, but the results are exactly the same as those obtained above.

A factorial design estimates 'main effects' and 'interaction effects'. The main effect of an intervention is the average effect of the intervention in comparison to the effects

observed under different combinations of the other interventions. For example, in a simple 2-by-2 design, the main effect of A is the average of the effect of A in comparison to no intervention and of the effect of A and B combined in comparison to the effect of the intervention B. Box 2 uses a numerical example to illustrate how the main effects and the interaction effects of the factorial designs in Tables 6 and 7 are calculated. The key point is that, in a factorial design, the intervention is not only compared to a control group not receiving any intervention, but also to a series of groups receiving the other interventions in different combinations. The 'pure' control group used in a multi-arm trial corresponds to a small, often tiny, fraction of the sample of a factorial design. An interaction effect in a factorial design is the effect of a factor that varies at different levels of another factor. The ability to estimate interaction effects is what makes factorial trials particularly appealing in the evaluation of complex interventions.

Factorial designs allow including as many factors and interactions as required with relatively small sample sizes. However, researchers hoping to use factorial designs to evaluate the effectiveness of different interventions and of their interactions may be disappointed. As mentioned, factorial designs do not estimate the effect of interventions against a control group not receiving any intervention. In a factorial design, the estimated effect is the average of the effect of the intervention versus the effects of other interventions and of their combinations, an object that is not easy to interpret. Main effects can be estimated with sufficient statistical power in a factorial experiment. However, the statistical power of the comparison between a particular intervention and the absence of any intervention will be small. This means that factorial designs will not normally be statistically powered to estimate the effectiveness of specific interventions against a control group without intervention, which is the focus of all experimental evaluation. What, then, is the use of factorial designs?

Unlike randomised trials, factorial designs are exploratory in nature. Issues of statistical power are less prominent than in standard randomised experiments, because the goal of factorial designs is not to estimate the population-level effectiveness of an intervention against a pure control group. The goal of a factorial design is to identify promising interventions and their combinations. They are appropriate in cases where there is little knowledge of the phenomenon at hand, and where there is uncertainty regarding what particular interventions are effective and in what combinations.

A factorial design is not to be interpreted as a one-off exercise, but as an element of a process of scientific discovery. When there is uncertainty and limited knowledge about the effectiveness of interventions, assessing them one at a time using a randomised trial is not an efficient strategy. Factorial designs allow for the identification of the

interventions that appear more promising. In addition, and importantly for complex interventions, they allow the discovery of relevant interaction effects between different interventions. For example, large interaction effects are worth pursuing and are also more likely to be detected with a given sample size. In a second stage, another factorial experiment can be conducted using only the relevant factors and interactions of factors identified at the first stage. Eventually, an RCT can be carried out to test the effectiveness of the interventions identified in the second stage.

Factorial designs are rarely used in evaluation. We reviewed the use of factorial designs in the evaluation of development interventions and found 27 studies. However, nearly all the studies consisted of 2-by-2 factorials. With just one exception, they were all designed as multi-arm trials rather than as factorial designs. In some cases, the design was slightly more complex than a simple 2-by-2 factorial. For example, some studies employed a 'cross-cutting' design in which the researchers randomised intervention A and then further randomised intervention B within the arm receiving treatment A. In some cases, there were more than two levels per factor. For example, a study promoting the use of condoms in motels in Nicaragua offered condoms on request, in room, or handed them out to customers, with or without health education material (Egger *et al.*, 2000). These studies are extremely simple and do not try to assess the effects of different interactions between interventions. Indeed, these studies are conducted within the traditional experimental approach of estimating the population effects of interventions using an appropriate sample size. Multi-arm designs employing many interventions and factors, as in factorial designs, are in fact impossible to conduct because they would require unattainably large samples.

As mentioned, only one of the studies we reviewed employed a true factorial design. Lachman *et al.* (2019) designed a 2-by-3 factorial trial to test the impact of parenting interventions on the behaviours of children and parents—such as child aggressive behaviour, dysfunctional parenting, and positive parenting—in three countries (Macedonia, Moldova, and Romania). The factorial included three interventions at two levels: a parenting session component (*short*: five sessions, or *long*: 10 sessions); a parenting engagement component (*basic*: childcare and transportation support and a simple snack or *enhanced*: food parcels, a raffle prize, rewards for attending 80% of the programme, and text reminders); and a facilitator supervision component (*heavy*: five supervision sessions with a trained coach, or *light*: supervision from trained coaches only upon request). With three factors and two levels, there were eight combinations of interventions and their interactions. In each country, two clusters (with 18 participants per cluster) were assigned to each intervention group. Across the three countries, the study included a total of 48 clusters and 864 participants.

The goal of the study by Lachman et al. is not assessing the effectiveness of the three interventions. The study follows a multiphase optimisation strategy over three phases (preparation, optimisation, and evaluation), which has been used in the evaluation of other interventions, namely smoking cessation, drug abuse prevention, HIV prevention, and adult weight loss. Concretely, in the first stage the factorial trial identifies those interventions and interactions of interventions that appear more effective. In the second stage, the researchers conduct a randomised evaluation only including those interventions and combinations that were found to be promising in the factorial trial. The experiment is currently underway and the results are not yet available.

We reviewed the remaining 2-by-2 factorial multi-arm trials in order to assess the significance of the interaction effects. Interaction effects are the simplest form of complexity, denoting the presence of synergies between intervention components. It is of interest to observe how many studies found significant interaction effects. One of the studies explicitly acknowledged being unpowered to detect the statistical significance of interactions and did not report any result (Burke, Chen and Brown, 2018). No other study mentioned a lack of statistical power, although this is a common problem of studies testing interaction effects. Six studies reported the results of main effects, but not the results of the interaction effects. This is a strange choice, considering the goal of factorial design is precisely to test the presence of interactions. However, these six studies did not find statistically significant main effects, and we can probably assume that interactions were not tested or were not found to be statistically significant. Thirteen studies did not find statistically significant interactions. In some of these studies, the researchers detected some interaction effects for some of the outcomes. However, considering that many outcomes were considered at the same time and that some of the interaction effects were negative, it is possible that these few statistically significant results were the result of random variation.

Three studies found clear statistically significant interactions. A study providing cash incentives and cognitive behavioural therapy to adult men involved in crime and violence in Liberia (Blattman, Jamison and Sheridan, 2017), found that cash alone and therapy alone reduced crime only in the short term, but the two interventions together decreased crime substantially and for a longer period. A trial providing school grants and teacher incentives in Tanzania (Mbiti *et al.*, 2019) found no impacts of school grants on test scores, and some positive effects of teacher incentives. However, the study found significant positive effects when grants and teacher incentives were simultaneously provided. A study of community-based interventions to reduce blood pressure in Pakistan (Jafar *et al.*, 2009), found that the combination of health education

by community health workers and training of GPs on hypertension management had a larger impact than when the interventions were independently implemented.

Two studies found some statistically significant interactions for some of the outcomes considered. A study providing peri- and post-partum infant feeding to prevent the transmission of HIV in Botswana (Thior *et al.*, 2006), found some interaction effects for some of the outcomes. A study by Yousafzai *et al.* (2015) provided responsive stimulation and nutrition supplements to caregivers to assess the impact on parenting skills, emotional distress, and caregiving capacity, and found some statistically significant interactions for some of the outcomes considered.

The inability to find statistically significant interactions is not limited to multi-arm trials conducted in international development. Mdege *et al.* (2014) conducted a systematic review of 2-by-2 factorial trials in clinical research and healthcare settings, and found 29 studies. Of these, 21 assessed the presence of interactions, but only four found statistically significant interaction effects. Li *et al.* (2006) conducted a meta-analysis of 113 factorial experiments in manufacturing and found that only 11% of the 2-factor interactions considered, only 6.8% of the 3-factor interactions, and only 2.8% of the 4-factor interactions were statistically significant.

Li *et al.* (2006) discuss their findings as evidence of three common regularities of factorial trials:

1. effect sparsity: the number of important effects in a factorial design is generally small and most interventions turn out to be ineffective;
2. hierarchical ordering: main effects are on average larger than 2-factor interactions, 2-factor interactions are on average larger than 3-factor interactions, and so forth; and
3. effect heredity: effective interactions are more likely to be found when their parent factors were also effective.

These findings seem to suggest that synergistic effects are rare and that interventions are not overly complex, in the sense that combinations of interventions rarely produce synergistic effects. Generalising this conclusion to development interventions, however, would not be appropriate, because all the multi-arm studies found by our review included just two factors and one interaction and were therefore not designed to explore the complexity of the interventions considered. In addition, even if interactions are not common, they can be of great policy interest. The fact that few interactions are normally found should not deter researchers from exploring their presence, particularly when there are good theoretical reasons for expecting them.

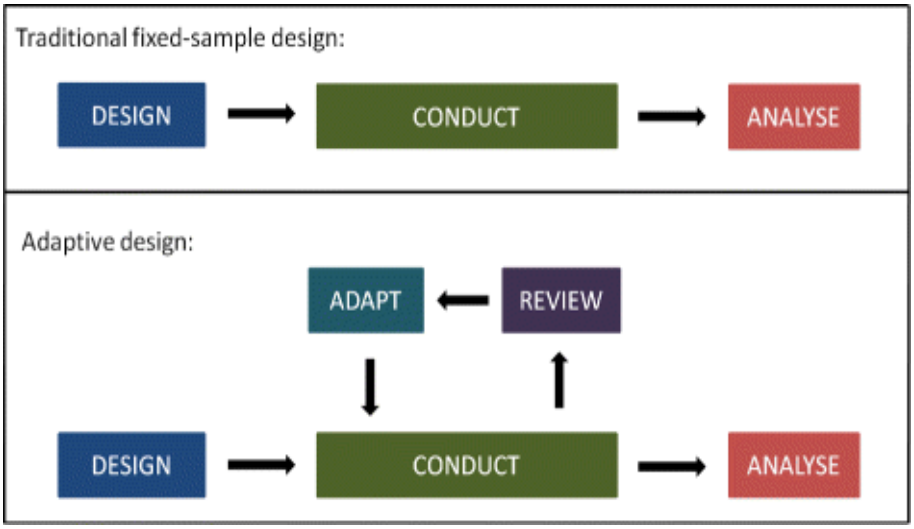
5.3 Adaptive trials

Adaptive trials are not new. They were the subject of intense discussion in the biostatistics literature of the 1960s and of the 1970s. The discussion was largely theoretical; even today empirical applications are few, and mostly concentrated in the drug industry. In this section, we provide a brief description of adaptive trials, we review adaptive trials conducted in international development, and we conclude with some comments on their potential contribution to the evaluation of complex interventions.

Our description of adaptive trials is based on the United States Food and Drug Administration guidelines for conducting adaptive clinical trials (2019) and on the introductions by Pallmann *et al.* (2018) and by Thorlund *et al.* (2018). An adaptive trial is a randomised experiment that allows for changes in the study design during implementation based on the data collected in the early stages of the study. The diagram below from Pallmann *et al.* (2018) illustrates the key difference between a randomised trial and an adaptive trial. In a randomised trial, the researchers design the experiment, they randomly allocate subject to different treatment groups and collect relevant data, and finally they analyse the results. In an adaptive trial, there is a feedback loop operating at the implementation stage. After collecting the first set of data, for example the baseline data, the researchers change the study design based on their preliminary analysis of the data.

The design of a trial can change for several reasons that fall under the categories of statistical efficiency, ethical acceptability, and improved understanding. First, adaptive trials are statistically more efficient. Researchers can use preliminary results to change the allocation of subjects across arms—or to change the number of arms—without compromising statistical power, thus saving resources. Second, adaptive trials are more likely to be accepted by patients and by public officers for ethical reasons. In an adaptive trial, researchers can discontinue ineffective treatments and expand the coverage of effective ones, thus preventing the wastage of resources and the administration of unnecessary procedures. Third, adaptive trial can be used to learn to understand how the interventions work along the causal chain or for specific subjects. In an adaptive trial, researchers can reallocate the interventions to specific groups or subjects with specific characteristics in order to test specific hypotheses based on a prespecified theoretical model, and thus obtain a deeper understanding of how the interventions work.

Figure 4: Design process of randomised trials and adaptive trials



Source: Pallmann *et al.* (2018), p. 2

Changes in an adaptive trial can occur in various ways, including:

- *group sequential designs*: researchers can stop some treatments, or the entire trial, when a midterm interim analysis shows these are not effective—the experiment is said to stop for 'futility';
- *adaptation to sample size*: researchers re-estimate the statistical power of the study using baseline data and change the sample size;
- *adaptive enrichment*: researchers use the interim data to identify subgroups in the population for whom the treatment is particularly effective, and refocus the attention of the study onto this particular group;
- *adaptation to treatment arm selection*: researchers remove or add treatment arms based on the observed effectiveness of the interventions—this is often used in 'dose-finding' trials, where different subjects are allocated to different doses of the treatment;
- *adaptation to subject allocation*: the researchers change the proportion of subjects assigned to each treatment arm based on the preliminary effectiveness of the intervention (these are also known as 'arm bandit' experiments);
- *adaptive hypotheses*: researchers change the research questions, for example switching from testing single to testing multiple hypotheses;
- *adaptation to endpoint*: researchers simply change the date of termination of the trial; and
- *multiple adaptation*: researchers introduce more than one of the changes above at the same time.

Most empirical applications of adaptive trials have been in drug testing. Bothwell *et al.* (2018) reviewed the use of adaptive designs in clinical trials and found 142 studies, of

which 85% tested specific drugs. They found that 21% of adaptive trials where sequential, including prespecified options for changing treatment arms, sample size, and study endpoint. Another 20% of studies collected biomarker data to select subgroups for a more focused analysis. Dose-finding (testing different dosages of the same treatment) was also common (16%). Finally, 9% of studies adopted a 'pick-the-winner' approach by using the interim data to select the best treatment for continuing the study.

Outside clinical research, adaptive trials have found very few applications. This is partly because of the lack of expertise and familiarity with this method among researchers and funders. Adaptive trials have also a number of limitations both on the practical and on the statistical side. On the practical side, the implementation of an adaptive trial is more demanding than conducting a traditional trial. First, changes to the design have to be planned in advance in order to avoid cherry-picking the results, which in turn requires a deep understanding of treatments and contexts. Second, changes require the timely collection and analysis of interim data. Third and finally, changes such as the introduction of new treatments and hypotheses may lead to results that are not easy to analyse or interpret.

On the statistical side, it is claimed that adaptive trials can produce misleading results. In particular, they may lead to the rejection or acceptance of wrong hypotheses. This occurs because adaptive trials often operate by changing the allocation of subjects to treatments. However, the selection of observationally successful treatments will sometime lead to selecting treatments that are effective only by chance. Similarly, effective treatments can be discontinued because by chance they appear to be ineffective. In summary, adaptive trial can produce two types of error: the acceptance of ineffective treatments (false positives) and the rejection of effective ones (false negatives).

In our review, we were only able to find two studies that employed an adaptive design in the evaluation of a development intervention. Caria *et al.* (2020) conducted an adaptive trial to evaluate a labour market intervention in Jordan. The intervention promoted job-seeking among Syrian refugees. Subjects were randomly allocated to three intervention groups and to a control group. The first group received a small and unconditional cash transfer. The second group received personal coaching to prepare for a job interview. The third group was 'nudged' by providing a planning calendar for job search and an instruction video on the use of the calendar. The control group simply received an information flyer. Subjects were interviewed after six weeks and again after two and four months. After the first follow-up, the treatments were reassigned to the

most effective treatment based on observed employment outcomes. The reassignment employed a 'Thompson algorithm', which changes the probability of assignment to a treatment group based on the effectiveness of the treatments. In the end, the three interventions were moderately and equally successful in promoting employment, in such a way that the adaptive reassignment mechanisms produced limited benefit.

Kasy *et al.* (2021) conducted an adaptive trial of an experiment in the provision of agriculture extension services in Odisha, India. The project contacted farmers by phone and administered a short questionnaire to enrol farmers in the programme. The study focused on enrolment in the programme as a measure of the success of the intervention. Six different types of call were trialled: a call at 10:00 am; a call at 10:00 am preceded by text message one hour ahead; a call at 10:00 am preceded by a text message 24 hours ahead; a call at 6:30 pm; a call at 6:30 pm preceded by text message one hour ahead; and a call at 6:30 pm preceded by a text message 24 hours ahead. A first batch of calls were made to 600 farmers randomising the six interventions. The researchers used the results of the calls to estimate the probability of completing the call in each intervention group using a modification of the Thompson algorithm. A further 17 waves assigned farmers to the same six treatments based on the estimated probability of success of the different calls in the previous waves. As in the previous study, the effectiveness of the calls turned out to be similar across treatment groups. The experiment nevertheless showed a marginal improvement in conducting calls at 10:00 am with a text message sent one hour ahead, and the allocation of subjects to the most effective treatments increased overall farmers' enrolment rates.

Both studies were of the 'adaptation to subject allocation' type and adopted a similar design. The sequential selection of the most effective interventions appears to be the most popular application of this approach. Although the two empirical applications reviewed tested simple treatments and produced results of limited policy interest, they are nevertheless good illustrations of the potential of this experimental approach. Adaptive trials are useful to evaluate complex interventions consisting of multiple components that are very similar and for which there is limited knowledge regarding their relative effectiveness. For example, they could be employed in the evaluation of adaptive interventions that change their activities along implementation. In principle, the same approach could be employed to assess the effectiveness of interactions between components and hence to shed light on the effectiveness of even more complex interventions. As discussed above, adaptive designs can also improve randomised trials in other ways: by increasing statistical power, by focusing on specific population groups, and by reformulating research hypotheses.

5.4 QCA

QCA is a research method that integrates qualitative and quantitative approaches. It was largely developed by Ragin and co-authors in a series of books (Ragin, 1987, 2008; Lieberson and Ragin, 2001). This method adopts a configurational understanding of causality and is sometimes depicted as a paradigm shift in comparison to traditional methods of causal inference. QCA is grounded in the notion that effects are produced by multiple causes and, as will become clearer in what follows, it bears some similarity to factorial experiments. Although the method does not build a counterfactual, the application of sufficient and necessary conditions for the identification of causal factors could be interpreted as implicit applications of counterfactuals.[4] One difference with factorial designs is that the set of combinations of causes considered by QCA is not built experimentally, but simply observed by employing the available data. Nor does QCA address the selection bias implicit in dealing with a subset of the potential combination of causes, as quasi-experimental studies normally do. It is therefore a comparatively weak method of causal inference.

The very first study employing QCA appeared in 1984, but the approach became popular towards the late 2000s when QCA studies started to increase at an exponential rate (Roig-Tierno, Gonzalez-Cruz and Llopis-Martinez, 2017). QCA is predominantly used today in political science, in sociology, and in business and management studies (Rihoux and Marx, 2013). As a research method, it has various goals(Rihoux and Marx, 2013): summarising qualitative data in a systematic way; testing theories and new hypotheses; and developing new hypotheses or theories. However, it recently gained acceptance in the evaluation literature and has been used to evaluate interventions or to analyse sets of interventions in systematic reviews (Thomas, O'Mara-Eves and Brunton, 2014). In this section, we first describe the main characteristics of this approach. We then discuss the QCA studies of development interventions that we found in our review, and we conclude with some comments on the use of QCA in the evaluation of complex interventions.

QCA is a small-N method.[5] Although there are studies employing hundreds or even thousands of observations, the method was devised in the tradition of comparative case studies to deal with a limited number of observations. It requires a deep knowledge of the cases considered and an iterative process of analysis between

[4] For example, this is the case when QCA qualifies the factor B as 'redundant', when the causal combinations ABC and AC are associated with the same effect.

[5] Our description of QCA is based on the textbook by Schneider and Wagemann (2012). We also used the notes to the QCA course offered on Coursera by Fadi Hirzalla of Erasmus University (www.coursera.org/learn/qualitative-comparative-analysis).

quantitative and qualitative observations. A review of QCA studies found that, in 60% of cases, between 10 and 50 observations were considered (Rihoux and Marx, 2013). The method is therefore well suited to analyse differences occurring between aggregates such as countries, regions, and organisations.

QCA adopts a configurational notion of causality. In QCA, events are the result of different configurations of multiple factors, or 'conditions'. Unlike typical quantitative research, QCA does not aim at isolating the effects of single causes after controlling for the impact of other factors. On the contrary, it considers all factors simultaneously, with the aim of identifying 'recipes' or combinations of factors that are associated with the effects.

The concept of multicausality used in QCA can be traced back to the work of Mackie (1965). Mackie discussed how traditional notions of 'causes' mostly refer to INUS conditions, that is elements of combinations of causes that are not able to produce effects on their own but are nevertheless necessary to the working of the combination of causes. In turn, each combination of causes is unnecessary to produce the effect, as other combinations may give rise to the same effect, but it is nevertheless sufficient to produce the effect. Rothman and Greenland (2005) discussed the same concept of multicausality in the form of 'causal pies'. QCA also employs a specialised terminology: multicausality is called 'conjunctural causality' or 'complex causality'; causal pies are called 'configurations' or 'conditions'; and the fact that different configurations may produce the same outcome is called 'equifinality'.

QCA is a case-based approach, and the first step of a sound QCA analysis is the identification of relevant cases and causal conditions. The method requires a deep knowledge of single cases, and it normally starts with the collection of data and the analysis of individual cases. Once cases and likely combinations of causes have been identified, researchers proceed to build a 'truth table', which lists all possible configurations of inputs associated with an outcome. A truth table, of which Table 8 is an example, is analogous to a standard form table of a factorial design (see the similarity between Table 8 and Table 7). There are however some differences. QCA employs observational data; therefore, some of the configurations of inputs are not observed. For example, the last three columns in Table 8 report the number of configurations with a positive outcome, the number of configurations with a negative outcome, and the type or number of cases found. Notice that, for some combinations of events, no outcomes are observed. These cases are called 'remainders' by QCA and are a symptom that reality offers 'limited diversity'. Another key difference between a QCA truth table and a standard form table of a factorial trial is the presence of 'contradictory'

cases in QCA. In QCA, when the same configurations result in different outcomes, it is called a 'contradiction'. QCA does not allow contradictions, which have to be resolved before analysing the data. Approaches to address contradictions and remainders have an *ad hoc* flavour,[6] but in some applications these are addressed through a rigorous iterative process of quantitative and qualitative analysis—for example, through interviews with key informants or participatory processes with key stakeholders.

Table 8: Example of a QCA truth table

Configurations	Intervention A	Intervention B	Intervention C	Outcome Y = 1	Outcome Y = 0	Cases
1	–	–	–	0	2	2
2	+	–	–	2	0	2
3	–	+	–	1	0	1
4	+	+	–	4	0	4
5	–	–	+	1	2	Contradiction
6	+	–	+	–	–	Remainder
7	–	+	+	3	0	3
8	+	+	+	–	–	Remainder

The data in the truth table are then simplified and summarised through a process of 'minimisation'. The minimisation process operates much like an Occam's razor by eliminating redundant configurations. For example, supposing both combinations AB and ABC are sufficient causes of the outcome, only the combination AB is retained while

[6] Three strategies are recommended to solve contradictions. The first consists of adding another configuration based on the assumption that if two combinations result in the same outcome, it must be because some other relevant factors were not taken into account. A second approach consists of redefining the outcome. For example, suppose that success of the outcome depends on a metric above a 0.5 threshold, which results in a contradiction. Changing the threshold may result in reclassifying the contradictory combinations as being either failures or successes. Third and finally, as a last resort, one or more of the contradictory cases can be removed from the data. Three potential solutions are proposed for remainders. The first consists of ignoring combinations for which no outcomes are observed. This is often not considered satisfactory, however, and a 'standard analysis' is proposed that leads to two other possible solutions: a 'parsimonious' solution or to an 'intermediate' one. Basically, the standard analysis simulates the outcomes for the unobserved cases and makes assumptions about the outcomes. Theory can be employed to justify the choice of a particular direction of outcomes for some combinations. The parsimonious and the intermediate solutions produce different results, which are based on different assumptions about the unobserved distribution of outcomes for some conditions. The exercise is normally performed by specialised software, and the intermediate solution is often recommended as providing a more easily interpretable number and type of solutions.

the combination ABC is eliminated, because by comparison it is apparent that the element C is redundant to the occurrence of the outcome. The process of minimisation is more complicated than described because it applies to all pairwise comparisons and it is normally produced by software, but the principles of logical consistency and of redundancy are the same.

Minimisation produces a number of configurations that are associated with the outcome. For example, the typical result of a QCA analysis may look like this:

$$AB+BC+{\sim}AC \rightarrow Y$$

which should be interpreted in the following way. The success of the outcome Y is equally produced by the presence of three different configurations of inputs: A and B, B and C, and C but not A. These are different 'recipes' associated with the outcome.

For simplicity, we limited our discussion to binary inputs and outputs, or 'crisp sets' in the QCA terminology. Crisp sets are easier to illustrate and also to interpret, but researchers often use 'fuzzy sets', which employ continuous values for both the inputs and the outputs. Researchers are often interested in phenomena and causes that vary by degree rather than dichotomous causes. The possibility of using continuous variables greatly increases the range of applicability of QCA to various types of problems and data. Fuzzy sets also allow partial set membership, whereby cases are not forced to be full members of a specific set. The use of fuzzy sets raises a number of challenges in the minimisation process. These can nevertheless be addressed by specialised software, or through participatory processes with key stakeholders.

QCA has a number of limitations. First, it relies on observational data and can only be conducted retrospectively. Second, as discussed in relation to contradictions and remainders in Footnote 6 QCA makes a number of oversimplifying assumptions regarding the observed and unobserved data that are rarely supported by robustness analyses. Third, it is best suited for small-N studies and does not perform well when there are many observations and when configurations are composed of many inputs. Often the number of attributes considered is fewer than five, because as their number increases the analysis requires more and more observations. In addition, with many attributes, the successful configurations identified become difficult to interpret. Finally, the robustness of the conclusions is also driven by the size of the sample, which by construction tends to be small in relation to the attributes considered.

Despite these limitations, QCA can nevertheless be fruitfully employed in the evaluation of some types of intervention. In our review, we identified eight impact evaluations of

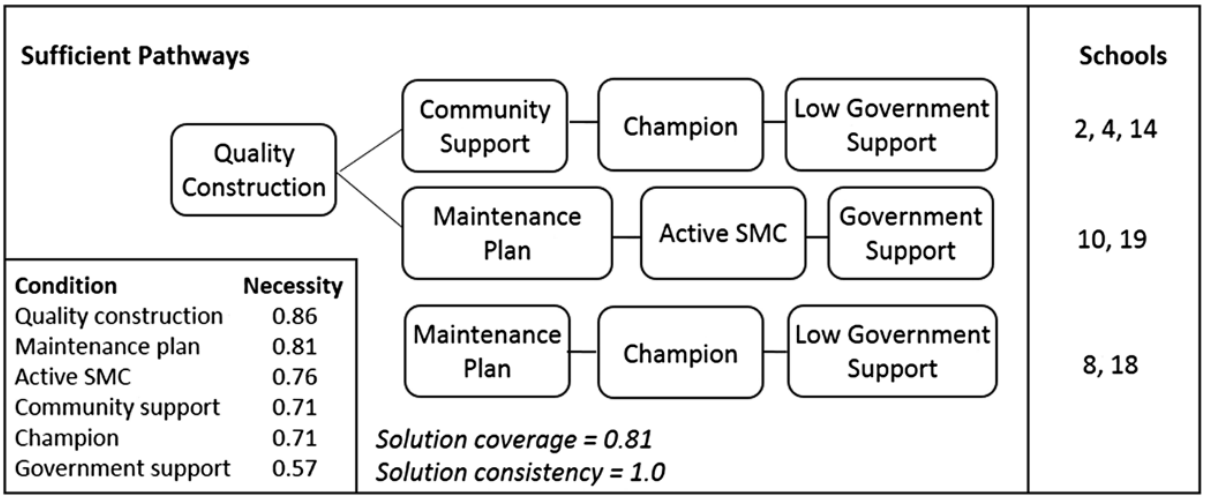
development interventions using QCA. The studies evaluated projects from various sectors including public health, urban infrastructure, irrigation, sanitation, and humanitarian assistance.

With just one exception, the studies identified evaluated the interaction of the project with the characteristics of the environment, rather than the interactions between project activities. An *et al.* (2017) assessed the impact of 11 'capacity development factors'—such as the availability of human resources, financial capacity, and the implementing agency—in conjunction with an urban infrastructure development in India. Chatterley *et al.* (2014) evaluated the impact of a series of environmental conditions on the effectiveness of an intervention providing hygiene and sanitation facilities to schools in Bangladesh. The conditions considered were high-quality construction, community support for maintenance, government support for maintenance, an active school management committee, the presence of a maintenance plan for sanitation, and the presence of a sanitation champion. Davis *et al.* (2019) evaluated the impact of nine project and environmental conditions, such as the presence in the community of behavioural change education, the absence of government barriers, and the availability of technical support for maintenance, on a sanitation intervention. Lam *et al.* (2010) investigated the conditions associated with the effectiveness of a large-scale irrigation project in Nepal, focusing in particular on the availability of continuous assistance, the presence of written rules, the imposition of fines, leadership, and collective action. Opdyke *et al.* (2019) analysed how eight different forms of community participation—such as involvement in needs assessment, unpaid labour contributions, and approval from the local municipality—influenced the effectiveness of a post-disaster shelter project in the Philippines. Pattyn *et al.* (2019) investigated 10 conditions, such as experience and education, and characteristics of the media employed, for example level of coverage, salience of information, and regional focus, on a media support intervention in Kenya and Tanzania.

Only one study assessed configurations of different project characteristics. Breuer *et al.* (2018) evaluated the impact of a multicomponent mental health intervention in Nepal. The intervention included a package of activities to promote service utilisation, such as the supply of medication, supervision, facility staff training, facility psychosocial interventions, referral to tertiary care, community awareness, and stigma reduction. These activities were implemented to a different degree by different facilities, allowing the researchers to explore what combinations were more effective in increasing service utilisation.

The results of the studies reviewed were reported through narratives of how the different configurations affect the outcome, and are too long to be reproduced here. Sometimes, the results are illustrated with diagrams, like the one in Figure 5 from Chatterley *et al.* (2014), which shows the configurations of conditions associated with the success of management of school sanitation in Bangladesh.

Figure 5: QCA pathways to well-managed school sanitation Bangladesh



Source: Chatterley *et al.* (2014), p. 9

In some cases, the conditions are represented in algebraic form. For example, Lam *et al.* (2010) summarise their results in the following way:

$$\mathrm{T} = \mathrm{CRf}(\mathrm{a} + \mathrm{L})$$

which means that the simultaneous presence of collective action (C), written rules (R), and the absence of fines (f) are necessary conditions for the irrigation programmes they investigated in Nepal to sustain agricultural productivity. The necessary elements, however, are insufficient to produces productivity increases. Productivity increases will occur only if either one of two additional conditions are present: the absence of external assistance (a) or the presence of leadership (L).

These examples show that QCA can be a powerful tool in shedding light on the combinations of activities and environmental characteristics associated with project success. In the majority of cases reviewed, the researchers investigated the interactions between the projects and the characteristics of the population, the circumstances of implementation, and other contextual factors. However, nothing prevents researchers from considering interactions between intervention components, as in the study by Breuer *et al.* (2018). QCA can therefore potentially be used in the evaluation of multicomponent interventions. In addition, given its reliance on a small number of cases, it appears particularly suited to the evaluation of portfolio interventions, when

there are many interventions in the same sector in the same country or across countries. This is particularly true for portfolio interventions seeking to change complex and hard-to-measure outcomes, such as policy change (Pasanen *et al.*, 2019).

5.5 Synthetic control methods

Synthetic control is an evaluation method proposed in a series of articles by Abadie and co-authors (Abadie and Gardeazabal, 2003; Abadie, Diamond and Hainmueller, 2010). In a recent review of evaluation methods, it was described as 'the most important innovation in the policy evaluation literature in the last 15 years' (Athey and Imbens, 2017). The method was developed to evaluate the impact of large interventions and policies on single aggregate units such as countries, states, or regions. Standard quantitative evaluation methods rely on large number of observations and are unable to evaluate interventions implemented on a single unit, such as a country or organisation. Early applications assessed the impact of civil conflict on GDP in northern Spain (Abadie and Gardeazabal, 2003) and the impact of tobacco control legislation on cigarette consumption in California (Abadie, Diamond and Hainmueller, 2010). Successive applications have mostly evaluated state-level impacts of welfare and health system reforms, changes in legislation, taxation, and other wide-ranging policies (Bouttell *et al.*, 2018).

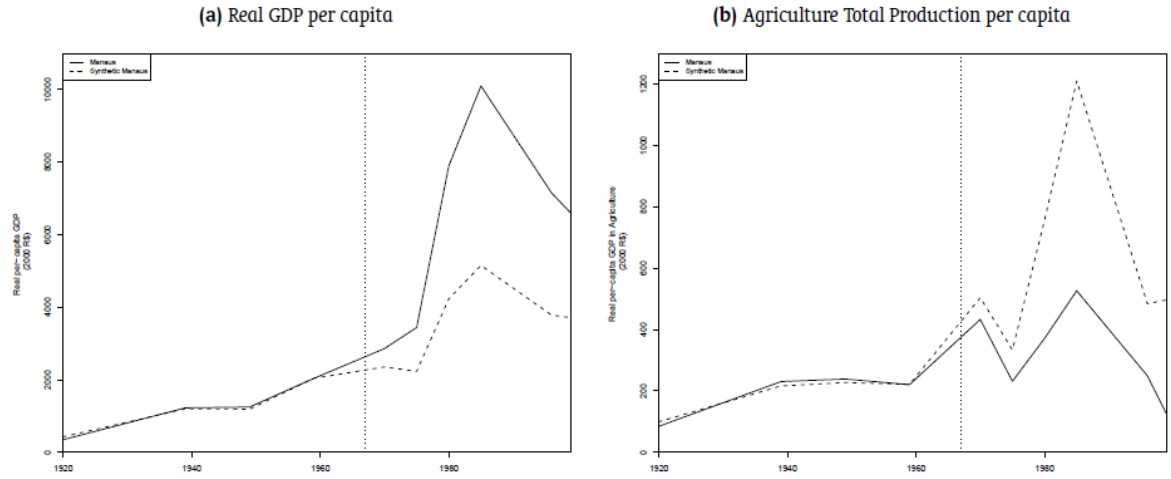
A typical synthetic control set-up involves an intervention in a country or state, and we have collected observations over a long period of time before and after the intervention for the treated unit as well as for a number of comparison units. The basic idea of synthetic control is that, while any one comparison unit is a biased comparison for the treatment unit, a combination of control units can provide an appropriate comparison for a project unit. Concretely, synthetic control uses pre-intervention information on the determinants of the outcome and assigns weights to different control units, giving more weight to those units that are more similar to the project unit during the pre-intervention period. The application of weights to the potential comparators generates a single 'synthetic control' unit, to which the project unit is compared. For example, in a study of the impact of the economic reunification of Germany, the GDP of Germany was compared to a 'synthetic' GDP consisting of a weighted average of the GDP of Austria, Japan, the Netherlands, Switzerland, and the United States (Abadie, Diamond and Hainmueller, 2015). Critically, the trend in the outcome before the intervention is included among the determinants, and is often the most important pre-intervention variable for the construction of the weights. Details of estimation procedures and of the statistical property of the estimator can be found in Abadie (2020). Software for

empirical implementation is available in different packages, such as Stata, R, and MATLAB.

Similarly to QCA, synthetic control builds a bridge between qualitative and quantitative approaches. A sound application of the approach requires a deep knowledge of the cases considered and of the determinants of the outcomes. Mechanical empirical applications of the method are unlikely to provide useful results. The method can be seen as a systematic approach to identifying appropriate comparison units in comparative case studies. The advantage in comparison to standard case studies analysis is the systematic approach to the selection of the comparator units and the ability to address causal inference.

A great advantage of synthetic control is that it displays the results of the analysis in charts that are self-explanatory and easily understandable. For example, the chart in Figure 6, from Possebom (2017), illustrates the impact of the establishment in 1967 of a Free Trade Zone in the city of Manaus (Brazil). The Free Trade Zone consisted of a set of new regulations on imports that provided large subsidies to manufacturers in the region. The researchers built a synthetic control for the region of Manaus using information from other regions of Brazil. The synthetic control should be interpreted as the counterfactual trend in GDP that would have occurred had Manaus not established a Free Trade Zone. The charts clearly show that the establishment of the Free Trade Zone had a large positive effect on GDP *per capita*, although it also had a large negative impact on agricultural production

Figure 6: Impact of the establishment of a Free Trade Zone in Manaus



Source: Possebom (2017), p. 226

Causal inference in synthetic control analysis relies on robustness tests, which assess the plausibility of the observed impacts against alternative plausible explanations.

These can be conducted in different ways. A typical application consists of considering 'placebo interventions'. The comparators are used in turn as intervention units, and impacts are estimated on the comparator units against the other comparators. The absence of observed effects on the units employed as comparators lends more credibility to the impact observed for the treatment unit. Another typical application consists of splitting the pre-intervention period in two halves and hypothesising the implementation of an intervention in the second half. The effect of this hypothetical intervention in the pre-treatment period is estimated, and again the absence of an observed effect gives more credibility to the impact observed on the treated unit. Robustness can also be explored by removing comparator units from the comparator pool one at a time. In this way, the researchers verify that the results are not sensitive to small changes in the composition of the comparison group and are not driven by any single comparator unit.

There are some difficulties in the interpretation of the results of a synthetic control analysis. A first case is when outcomes are volatile because of the operation of other factors, and it is difficult to separate impacts from noise. A second case is when the treatment unit or a comparator unit displays extraordinary values of the outcome, in the sense that they are so different from the values observed in the other units that a valid comparator cannot be found. To obviate these difficulties, the comparison units should be carefully selected, and units that are too different should be excluded. For example, in the Germany reunification study mentioned above, the researchers only used data from Organisation for Economic Co-operation and Development countries. In some cases, the similarity between units is obvious, but in other cases it requires additional investigation. Similarly, comparison units that have been affected by shocks should be excluded. For example, in the reunification example, Poland would not be a good comparator as it underwent a major economic shock at the same time as Germany. A third difficulty arises when the intervention produces spillover effects that radiate from the treatment unit to the comparator units—again a real possibility in the reunification example, particularly for Austria, the Netherlands, and Switzerland, given the weight of the German economy within Europe. This is not an unlikely case given the large scale of interventions evaluated by synthetic control methods. In the presence of a spillover effect, the estimated effects should be interpreted as lower or upper bounds of the true effects.

There are also some limitations that are specific to the evaluation of development interventions. First, the synthetic control method can only be employed retrospectively. Second, it works best with time series for treated and control units over a long period before and after the intervention. Effects that are observed over very short periods are

difficult to separate from naturally occurring random variations. For the same reason, the method works best when the impacts of the intervention are large. Unfortunately, development interventions often only produce small effects. The method will not provide useful results in the case of small interventions with effects that are difficult to detect.

Synthetic control is relatively underutilised in the evaluation literature. A recent review of empirical applications in the evaluations of public health interventions found only 38 studies using it, mostly from high-income countries (Bouttell *et al.*, 2018). Applications to specific interventions are rare, although one exception is the study by Gutierrez *et al.* (2016) assessing the impact of the Intensive Partnership for Effective Teaching—a large portfolio intervention funded by the Gates foundation—on academic achievements in selected American states. Another exception is the study by Bifulco *et al.* (2017), who used synthetic control to evaluate the impact of a district-wide scholarship scheme in Syracuse, New York.

Our review found eight studies assessing the impact of development interventions using synthetic control. The studies were conducted in a variety of sectors, including budget support, environmental policies, tax administration, and public health. Alavuotunki (2015) assessed the impact of donors' budget support on public health expenditure and neonatal mortality in 12 African countries. The study used 55 comparator countries and found that budget support had a significant impact on health expenditures in five countries. Positive effects were also found on neonatal mortality, although they were less conclusive. Barofsky *et al.* (2015) assessed the impact of a malaria control campaign in the Kigezi district in Uganda in 1959. They used birth cohort census data from 1929 to 1972 from 15 districts and found a significant impact of the campaign on schooling and labour market participation. Lepine *et al.* (2014) assessed the impact of the removal of user fees on health services in 54 districts of Zambia in 2006. They compared one district at a time to the remaining 18 districts between 1998 and 2006 and found an impact on health expenditure, but no significant effect on service utilisation. Quast *et al.* (2017) found a significant impact of the 2005 introduction of sex regulations (sex workers registration, HIV testing, and quarterly screening for sexually transmitted infections) on sexually transmitted diseases in Tijuana, Mexico. They used data from 1995 to 2012 and from 11 other Mexican states that did not implement any regulation as comparison units.

Chelwa *et al.* (2017) assessed the impact of tobacco control policies in South Africa (a large tax in 1994 and advertising in 2001). They used data from 1990 to 2001 from 24 countries with middle incomes and no tobacco policies, and found a significant impact

on cigarettes consumption. Sarr (2016) assessed the impact of establishing semiautonomous revenue authorities on tax administration. The study considered, one at a time, 20 countries that had established revenue authorities prior to 2000, using data covering the period from 1980 to 2010, and using 54 comparison countries. The results were mixed, as tax revenues increased in some cases, but decreased in others. Sills *et al.* (2015) estimated the impact of a local initiative to reduce deforestation in the municipality of Parguinas, Brazil, introduced in 2004. They compared the municipality to another 25 similar municipalities of the Amazon region over the period from 2002 to 2013, and found only marginal effects on deforestation rates.

Our short reviews highlight the great advantage of synthetic control methods: their ability to evaluate large-scale interventions and nationwide policies. Our sample of studies includes evaluations of budget support interventions, large-scale malaria campaigns, and the establishment of tax revenue authorities. These interventions are normally not evaluated, and surely not using quantitative methods. The approach appears to be best suited to evaluating large-scale portfolio interventions.

Our review also highlights some of the limitations of the approach, and its rather narrow range of application. First, the approach cannot always be implemented. Synthetic control requires sufficient data before and after the intervention in order to estimate project effects, but long time series are not always available. Even more serious are the issues of comparability with the control units. Large interventions are often implemented in specific countries or areas that are selected specifically for being more in need or poorest. As a result, it is often difficult to find suitable comparator areas. Second, the results of synthetic control analysis are sometimes difficult to interpret. Interventions are rarely implemented overnight and at a constant rate in such a way that delayed or variable effects are common.

Despite the above limitations, synthetic control appears to be a promising approach for evaluating large-scale interventions, particularly in view of its potential developments. Technical developments in statistical control approaches are a lively area of research. A similar, easy to implement, and easy to interpret approach is, for example, the causal impact of Brodersen *et al.* (2015). Causal impact was developed by researchers at Google and uses Bayesian structural time series analysis. It compares the treatment unit to a time series running over both the pre- and the post-intervention period. This requires that the series should not be affected by the intervention, and an obvious choice are the outcomes in the comparator units, or totally uncorrelated variable such as, for example, rainfall. The approach has been mostly used for marketing purposes, but nothing prevents its use in the analysis of development policies.

A second promising development is the synthetic difference-in-difference analysis of Arkhangelsky *et al.* (2018), which integrates the benefits of synthetic control and of difference-in-difference analysis. This approach estimates intervention effects by regression analysis (as in difference-in-difference), but only after weighting the observations (as in synthetic control analysis), the weights including adjustments for covariates as well as for time periods. Usually, the covariates consist of past values of the outcomes for the treatment and control units. Intuitively, the approach adjusts more heavily for differences in those time periods when these are more influential. This is normally equivalent to giving more weight to periods that are closer to the intervention period, thus preventing more distant periods from having a significant effect on the construction of the control group.

5.6 Agent-based modelling

Agent-based modelling is a popular approach for modelling complex systems. The approach has been used in various disciplines and it became a popular method in the social sciences for the simulations of interactions between individuals and their environment. Applications of agent-based modelling can be found today in the modelling of epidemics, financial transactions, migration, and waste management. Agent-based modelling, like system dynamics (discussed in the next section), is not a method of causal inference. It is nevertheless included in our review because models can be framed within counterfactual analyses simulating impacts with and without an intervention. In addition, the availability of data from actual interventions allows the validation of the model assumptions and testing their robustness. In other words, while agent-based models cannot be used alone to assess the impact of interventions, they have the potential to causally assess the effectiveness of interventions through the inclusion of counterfactuals and through the development of methods of validation and of robustness analysis.

We found six systematic reviews of agent-based modelling studies. Cassidy *et al.* (2019b) reviewed 11 agent-based modelling studies in the public health literature that simulated the behaviour of health systems. Kremmydas *et al.* (2018) found 32 agent-based modelling studies in the agricultural economics literature. Yang (2019) found 17 studies simulating behavioural interventions in the areas of diet management, drug abuse, smoking, and drinking. Ding *et al.* (2018) critically reviewed the agent-based modelling literature in construction waste management. Smith *et al.* (2018) found 90 articles using agent-based modelling to model malaria transmission. Nianogo *et al.* (2015) identified 22 studies using agent-based modelling to understand non-communicable diseases and

their public health risk factors. Finally, McAlpine *et al.* (2020) reviewed 28 articles using agent-based modelling in research on migration and modern slavery.

As can be appreciated by the large number of studies reported in these systematic reviews, agent-based modelling is not a new method, nor is it a method neglected by social scientists. However, agent-based modelling is little known in the evaluation and international development literature. In what follows, as in previous sections, we first describe the main characteristics of the approach; we then review its use in the evaluation of development interventions; and we conclude with some comments on its advantages and disadvantages for the evaluation of complex interventions.

Agent-based modelling can be characterised as a computer simulation of the behaviour of a social system (Marshall and Galea, 2015). In the simulations, agents' behaviours were guided by few behavioural rules and some parameters characterising the context environment. Utility maximisation and decision trees are common behavioural rules used in agent-based models. Concretely, researchers define a set of initial conditions at time zero, which define the state of the system at the beginning of the simulation; various characteristics of the agents and of the environment, such as demographic composition and geographic distances between agents; and a set of behavioural rules that characterise the agents' responses to other agents' actions and to the outcomes of the model.

The agents' responses to the behaviours of other agents are the key element of an agent-based model. It is this peculiar feature which allows the simulation of complex social systems. For example, complex patterns in the transmission of infections can be modelled by allowing individuals to interact with each other and to respond to the outcomes of the simulations. In this way, agent-based models sidestep the standard SUTVA, which is common to most statistical causal models (Imbens and Rubin, 2015). SUTVA assumes that the impact of an intervention on an individual is independent of the impact of the intervention on other individuals. It is easy to see that the SUTVA assumption, common to statistical model such as RCTs, rules out the presence of interactions among agents. Agent-based models make no such assumption. On the contrary, they formally theorise and simulate the transmission mechanism by which agent's behaviour is influenced by the behaviour of other agents. Interactions between agents produce non-linearities and discontinuities proper of complex systems.

Agent-based modelling has two great advantages in comparison to standard evaluation methods. First, as is true of other modelling approaches, it can evaluate hypothetical interventions through simulations. The method is unable to evaluate an intervention either retrospectively or prospectively, but the possibility of simulating the impact of

hypothetical interventions under different scenarios is appealing. Second, thanks to the ability to model agents' interactions, agent-based models can account for emergent phenomena. In other words, agent-based models are able to produce system-level outcomes that could not be predicted by observing the behaviour of single agents. The analysis of complex systems allows the identification of key trade-offs, which are important for policymaking as interventions that are beneficial to some sectors of society can be harmful to others in ways that rare not immediately obvious. They can also allow the identification of leverage points by identifying non-linear responses to inputs.

These advantages come at the cost of the limited plausibility of the results. Agent-based models are heavily dependent on the assumptions made. These assumptions are often simpler that those employed in economic models. These simplified models might be applicable to a range of social phenomena, but are unlikely to be valid to analyse interventions that produce changes, for example in norms and power relations (Mowles, 2014). As such, they find applications in agricultural or health interventions, but are less likely to be used in the evaluation of (for example) governance interventions.

Agent-based models are also dependent on the availability of data and of the calibration procedures used. Excessive parametric simplification, the absence of relevant data, a short time horizon, and underdeveloped sensitivity analyses are common limitations (Nianogo and Arah, 2015; Kremmydas, Athanasiadis and Rozakis, 2018; Cassidy *et al.*, 2019b). More generally, it has been observed that agent-based models have a difficult relationship with the data (Windrum, Fagiolo and Moneta, 2007). When using agent-based modelling, we must assume that all relevant causal mechanisms are included in the model and that they are correctly specified. This, however, needs to be demonstrated by testing the model with the data, but validation and sensitivity analysis are rather underdeveloped in agent-based modelling.

How can agent-based modelling be used in evaluation? In principle, once a model has been calibrated and validated with the data, it can be used to simulate and prospectively evaluate various hypothetical interventions. By construction, agent-based modelling evaluations are *ex-ante* and not *ex-post*. For inference, researchers can run the same model under different regimes—for example with and without the intervention—and record the outcomes. After many runs of the same model, they can compare the distribution of outcomes under the two regimes. Statistical tests can then be conducted to establish whether the difference in the outcomes under the two regimes is sufficiently extreme to consider the impact of the intervention as plausible.

In our review, we found 11 studies using agent-based modelling to evaluate the impact of development interventions. Six of these studies were conducted in agriculture. Belem *et al.* (2018) evaluated the impact of the adoption of new sorghum varieties in a 'virtual' village in Mali. The researchers considered two policies: general provision of fertiliser and equipment, and provision to farmer adopters. They simulated impacts over 40 years and found that farmers would adopt different varieties depending on the prevailing climatic conditions, which led them to recommend the development of high-yield seeds that are resistant to climate. Nguyen *et al.* (2019) assessed the viability of contract farming in the Vietnam Mekong delta. They built a model including complex adaptive trust formation among farmers, and simulated various degrees of contractor's commitment, prices, and availability of storage facilities. They concluded that contract farming was not cost effective. Nöldeke *et al.* (2020) simulated the impact of technology adoption among poor farmers in the Congo basin of Luapula province in Zambia. They simulated three interventions within a social network diffusion model: the selection of 'seed' farmers, the number of seed farmers, and the interaction between the two. They found that diffusion was maximised when farmers with the most connections were targeted, and that there were significant interaction effects between the two interventions. Widener *et al.* (2013) simulated the conversion of illicit poppy production in Afghanistan. They simulated the impact of insurgents' influence on farmers, the impact of blockades, and the impact of subsidies. They found that blockades were most effective in reducing poppy cultivation, and that small changes in insurgents' influence would require scaling up the interventions. Wossen *et al.* (2015) assessed the impact of agricultural credit and off-farm employment on poor farmers in the Upper East region of Northern Ghana. They concluded that climate and price variability have a disproportionate effect on poverty and livelihoods. Using a similar model, Schreinemachers *et al.* (2007) assessed the impact of loans, fertiliser, and improved maize on poverty of farmers in two villages of southeast Uganda. They modelled interactions between soil properties, yields, demography, diffusion of innovation, and price trends, and found that all interventions considered would reduce—but not eradicate—poverty.

Three studies were from the environmental sciences. An *et al.* (2020) evaluated the impact of payments for environmental services in the Guzhou province of China. They simulated three scenarios: the intervention as implemented, the removal of the intervention, and the doubling of the payments made to farmers. They predicted non-linear patterns of migration and land use over a period of 20 years. Andersen *et al.* (2014) assessed the impact of environmental interventions in a small forest village of Bolivia. They simulated the effects of public investments, green jobs, conservation payments, deforestation taxes, and incentives to prevent deforestation. The simulations

highlighted unavoidable trade-offs between reducing deforestation and improving living standards. Smajgl *et al.* (2015) evaluated the conversion of monoculture rubber in the Yunnan province of China. They simulated three scenarios: the use of payment for environmental services, the introduction of strict regulations, and the creation of off-farm employment in the tourism sector. They found that regulations were effective in reducing the rubber area but at the cost of impoverishing farmers. On the other hand, payments for environmental services had a perverse effect and tended to increase the rubber area.

Finally, we found two studies in the public health literature. Gharakhanlou *et al.* (2020) assessed the impact of malaria control interventions in the province of Sarbaz in southeast Iran. They expanded a standard epidemiological model of disease transmission by including the effects of interactions between agents, and between agents and the environment. They concluded that both indoor residual spraying and long-lasting insecticidal nets were effective when the coverage was medium or high, and that the two interventions were most effective when implemented together. Lemoine *et al.* (2016) assessed the impact of a 'bus rapid transit' system on walking in Bogota, Colombia. They simulated the impact of increasing the number of lanes and of expanding the number of bus stations across the city and found that expanding access to bus services increased commuters walking, but the relationship was non-linear and reached a plateau.

Our review highlights several of the advantages of agent-based model in evaluating complex interventions. All studies considered modelled complex systems. Some studies did so by modelling complex social networks (Belem, Bazile and Coulibaly, 2018; Gharakhanlou, Hooshangi and Helbich, 2020; Nöldeke, Winter and Grote, 2020), whereas others modelled complex agents' interactions with the natural environment (Schreinemachers, Berger and Aune, 2007; Smajgl *et al.*, 2015; An, Garvin and Hall, 2017). Through simulations, they prospectively evaluated interventions that could not be evaluated in field experiments such as the introduction of alternatives to illicit poppy production (Widener *et al.*, 2013). Two studies were able to evaluate interactions and synergistic effects between interventions (Gharakhanlou, Hooshangi and Helbich, 2020; Nöldeke, Winter and Grote, 2020). Models were used to make long-term predictions over up to 20 or 40 years (Belem, Bazile and Coulibaly, 2018).

Our review also highlights the main limitation of agent-based models. Only five studies conducted a validation of the model (Schreinemachers, Berger and Aune, 2007; Wossen and Berger, 2015; Lemoine *et al.*, 2016; An *et al.*, 2020; Gharakhanlou, Hooshangi and Helbich, 2020), and this was not always given the proper attention. Several studies

acknowledged an excessive simplification of the context analysed. Finally, some studies suggested that the model could be extended to include other context and agent characteristics (Wossen and Berger, 2015; Lemoine *et al.*, 2016; Nöldeke, Winter and Grote, 2020), but it is unclear whether this was done in practice.

These limitations notwithstanding, agent-based models appear well positioned in supporting the evaluation of intervention in complex systems characterised by emergent behaviours. This is particularly true when there is an interest in the systemic effect of interventions over a long period of time. Agent-based models are therefore potentially well suited to the evaluation of system-level interventions. However, they can also support the evaluation of multicomponent interventions, since they appear able to assess synergistic effects of interactions between activities. As micro-level studies, however, they are not designed for the evaluation of portfolio interventions. Their effective application to the evaluation of interventions nevertheless requires that these studies should be conducted alongside counterfactual studies, and that their results should be validated with real data and subjected to rigorous robustness testing.

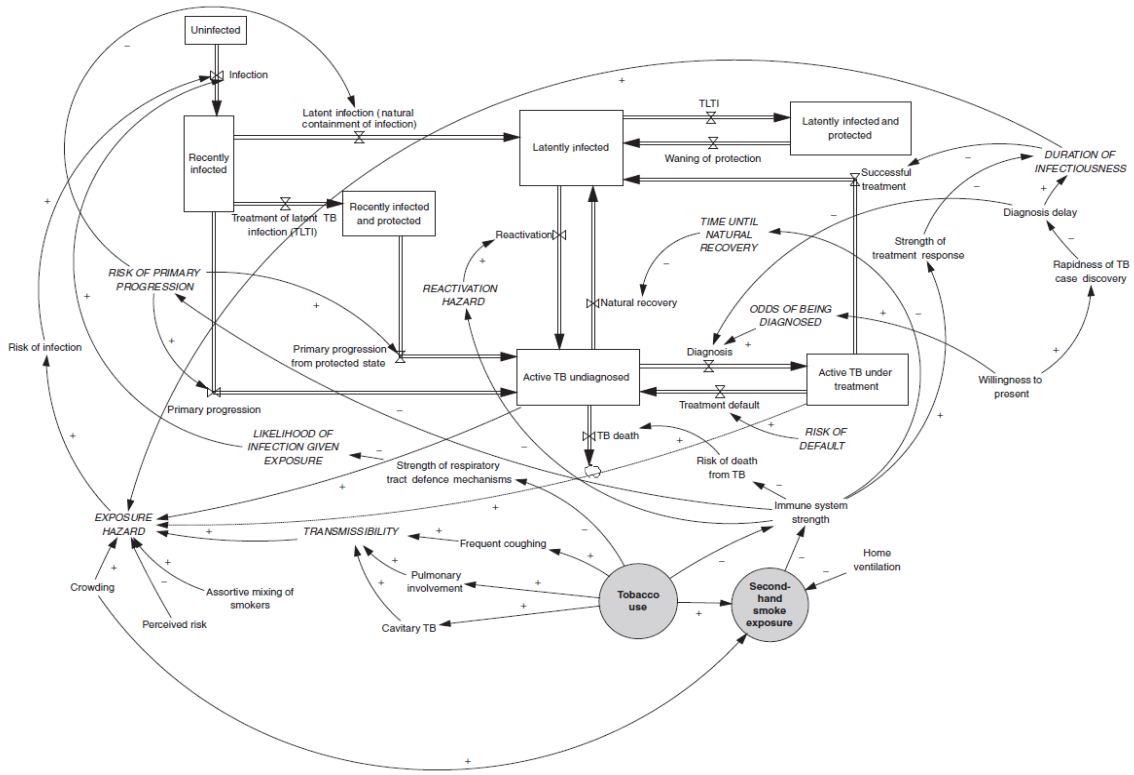
5.7 System dynamics

System dynamics is another popular approach for the modelling of complex systems. It was developed in the 1950s through the work of Forrester (1962) on industrial dynamics. It has been used in engineering and the social sciences, most notably in the influential *Limits to Growth* report commissioned by the Club of Rome in 1972. The approach has many followers from different disciplines, and there is an active System Dynamics Society (https://systemdynamics.org/), with an online bibliography listing more than 12,000 reference articles.

A fundamental tenet of system dynamics is that all elements of a system are interconnected (Sterman, 2002). Interconnectedness implies that every action produces a reaction. Changing one thing at a time, as it is normally done in experiments, is seen as a futile exercise because it ignores the chain of reactions produced by each action. Interconnectedness is a common modelling assumption, but system dynamics makes the additional assumption that connections are characterised by feedback loops. Some of these loops are negative, meaning that the reactions get smaller and smaller, and the system reaches some stability. Other feedback loops are positive meaning that reactions get larger and larger. Positive feedback loops are self-reinforcing mechanisms that generate exponential growth and instability. System dynamics is concerned with the evolution of actions and reactions over time and focuses on the long term. The approach invites researchers to adopt a system perspective by delineating the nature of feedbacks loops and by identifying the ramified consequences of different actions.

Concretely, system dynamics employs two tools: causal loop diagrams and feedback models. A causal loop diagram is a qualitative exercise that establishes the relationships between the elements of a system, and identifies positive and negative feedback loops. Loops generate stocks and flows of resources between the elements of the system. Causal loop diagrams may reach a high level of complexity. For example, the diagram in Figure 7 illustrates the relationship between tobacco consumption and tuberculosis (Hassmiller Lich *et al.*, 2010). Links between factors are represented by arrows. Arrows are positively or negatively signed depending on whether one factor causes an increase or a decrease on the linked factor. Unlike diagrams employed in much of statistical literature, in theories of change, and DAGs (directed acyclic graphs), the diagrams of system dynamics display feedback loops. Changes in factors produce changes that are feeding back, directly or indirectly, on the same factors. Flows of resources accumulate or decumulate stocks over time and are represented by boxes, while flows between stocks are represented by double-lined arrows. For example, in Figure 7, there is a stock of population infected by TB with in- and out-flows regulated by flow rates represented by the hourglass symbols. In the figure, the factors of interest—tobacco use and exposure—are represented in grey to map their interconnections within the system.

Figure 7: Causal loop diagram of tobacco use and TB



Source: Hassmiller Lich *et al.* (2010), p. 13

The second analytical tool used by system dynamics is simulation modelling, which formalises and quantifies the information included in causal loop diagrams.

Researchers link factors through mathematical relationships, set the levels of initial conditions for the stocks, and define the flow rates between stocks. To do so, they draw from observational data, expert opinion, or well-informed guesses. The model is then used to simulate different policy changes. Unlike agent-based models, which model interactions between units, system dynamic models take a macro view of a system, and consider interactions between large aggregates such as entire populations or populations subgroups rather than interactions between individuals.

System dynamics models are useful in various ways. First, they support the design of better policies by gaining an understanding of the long-term and unexpected consequences of interventions. Secondly, they help identify unexpected resistance in a system, which occurs when bottlenecks prevent policies from having an impact or lead to unexpected effects. Third and finally, by exploring feedback loops and self-reinforcing processes, they help identify leverage points and criticalities in the system.

One limitation of system dynamic models is the inability to validate the models with the data. Researchers can assess the robustness of system dynamic models by tweaking the functional and parametric assumptions of the models, and measuring the sensitivity of the results. Simulations however are difficult to validate because of the long-term horizon taken and the highly aggregate level of analysis adopted. A review of systemic dynamic models in public health (Cassidy *et al.*, 2019a) found that the difficulty to define the model parameters, the simplifications in their characterisation, the absence of required data, and the inability to simulate all relevant policy scenarios were common limitations.

Our review found eight system dynamic studies of development interventions. Five studies were in public health. Alonge *et al.* (2017) simulated the volume and quality of maternal and child health services provided by health facilities in a payment-for-performance scheme in Afghanistan. They simulated different hypotheses about gaming, levels of staff motivation, and bonus structures, and concluded that bonuses were likely to improve service quality, although this might be reversed in the presence of gaming. Feola *et al.* (2012) investigated the impact of personal protective equipment among potato farmers in La Hoya community in Colombia. They simulated seven different policies including safety labels, reduction of equipment cost, and social norm interventions, and concluded that changing farmers' behaviours is difficult, although some positive effects of normative pressure and of participatory process were obtained. Li *et al.* (2019) assessed the impact of factors influencing the provision of public health services by the Chinese Centre for Disease Control for the control of infectious diseases. The simulations suggested that increasing government investments and compensations

levels for health products would increase public health performance. Lich *et al.* (2010) built a conceptual framework illustrating the interactions between tobacco control interventions and tuberculosis (Figure 7). Semwanga *et al.* (2016) simulated the effectiveness of different interventions to reduce neonatal mortality in Uganda, such as increasing the demand for services, targeting health service delivery, and targeting the health of pregnant women. They concluded that a combination of free delivery kits and motorcycle coupons to take women to hospital during emergencies were the most effective interventions.

Three studies were in the water and waste management sector. Al-Khatib *et al.* (2016) estimated the amount of hazardous and general waste produced by three types of hospitals in the city of Nablus in Palestine to identify whether the private sector, the public sector, or charities were performing better in waste management. Borgomeo *et al.* (2018) simulated the effects of improving the reliability, operation, and maintenance of water infrastructure in coastal Bangladesh. They concluded that water infrastructure initiatives would allow farmers to escape from a poverty trap and would enable a virtuous circle of water improvements and assets accumulation. Prouty *et al.* (2018) assessed the effectiveness of various measures to increase the adoption of water waste resource recovery systems in Placencia, Belize. The results suggest that changing community behaviours is the most effective way to influence sustainable management of wastewater resources.

All the studies considered identified feedback loops between activities and contextual factors. In some cases, feedback loops were modelled with reference to existing theory. This led researchers to identify 'traps' or 'low equilibrium' points (Feola, Gallati and Binder, 2012; Alonge *et al.*, 2017; Borgomeo, Hall and Salehin, 2018). In all other cases, the feedbacks were not explicitly modelled, but the causal loop diagrams illustrated the complexity of the relationships between variables and how any change in the system will have ramifications.

Our review also highlights the limitations of system dynamics models. The models made big simplifying assumption and this was recognised by the authors, who stressed the need to incorporate more factors and activities in the future (Feola, Gallati and Binder, 2012; Li *et al.*, 2019) and to obtain better and additional data (Semwanga, Nakubulwa and Adam, 2016; Borgomeo, Hall and Salehin, 2018). Model validation was rarely conducted, and sometimes consisted of sharing the results with expert stakeholders in participatory exercises (Semwanga, Nakubulwa and Adam, 2016; Alonge *et al.*, 2017). More concerning is the fact that it is not entirely clear that system dynamic models can be validated at all. Many of the models made long-term predictions and were

dependent on such a large number of factors and initial conditions that a sensitivity analysis was difficult to carry out. This was recognised by the researchers, who rarely concluded with recommendations on policies and interventions. More often the researchers suggested that their results should be used to support the design of interventions (Alonge *et al.*, 2017), to inform discussions among stakeholders about policy options and potential courses of action (Semwanga, Nakubulwa and Adam, 2016; Prouty, Mohebbi and Zhang, 2018), or to inform the development of future models (Borgomeo, Hall and Salehin, 2018).

In summary, system dynamics is a useful approach for the evaluation of system-level interventions. The approach, however, is unlikely to be useful in the evaluation of multicomponent and portfolio interventions. The approach tackles the presence of complex phenomena directly by modelling feedback loops. However, its results are sensitive to model specification and rarely validated, such that they are mostly suggestive or speculative in nature. In other words, system dynamic models are not directly designed to support decision making but to build a better understanding of interventions and of the context in which they operate.

As was observed for agent-based modelling studies, system dynamics has been included in our review because it has the potential to evaluate hypothetical interventions. Its limitations should be weighed against their ability to model complex relationships and emergent phenomena. However, it must be noted that system dynamics studies in their current form, as found in the available literature, fall short of credibly assessing the impact of interventions. As agent-based modelling, they need to be integrated with methods of causal inference, either at the design stage through counterfactual studies including populations with and without the intervention, or at the analysis stage through rigorous validation against the data and sensitivity analysis. At a minimum, the predictions generated by the model should be tested in the field by a process of validation using real data from interventions.

6. Discussion

Our review did not cover all aspects of the evaluation of complex interventions. In some cases, the omissions were the result of a rational choice, because we believed a particular aspect not to be relevant. More often, the omissions were a pragmatic choice dictated by the limited scope of our review. In this section, we discuss the reasons for these omissions and their implications.

The study of complexity is normally associated with the analysis of non-linearities. By non-linearities, we mean the fact that two or more factors can simultaneously produce an effect that is larger than the sum of the individual parts, and that the impact of an input on an output is not proportional and varies with the existing levels of inputs and outputs. In the practice of international development, it seems that non-linearities can emerge in two ways: first, projects including more than one component may produce synergistic effects by interacting with each other; and second, changes produced by interventions may interact with societal or policy changes, thus generating similarly unexpected synergistic effects. We defined the first type of non-linearity as belonging to the complexity of the interventions, and the second as belonging to the complexity of the system. We then decided to focus on the first type of non-linearity. This is not because the second type of non-linearity is not common or interesting. We restricted our attention to complex interventions because the field is relatively unexplored and because this reduces the scope of the review. This choice has a number of implications for the types of project and of complexities considered. Some readers will notice the absence in our review of governance interventions and of complexities related to the context environment in which the interventions are implemented. The inclusion of challenges and methods to address complexities arising at the system level would require another review of similar length.

Evaluations of interventions can be carried out using a variety of approaches. In our review, we decided to consider only methods that address causal inference in a convincing way. We used a broad definition of causal methods that is not limited to counterfactual approaches, and which includes methods based on the notion of multicausality, such as QCA and factorial experiments, and modelling methods, the credibility of which depends on strong assumptions and much validation and robustness testing. We did not include methods with a more questionable causal approach, such as process tracing or contribution analysis, and we did not include qualitative approaches. This is not because we do not think these methods are useful. A thorough review of these methods and approaches would require a separate study, and similar reviews are already available in the literature.

In our review, we built a project typology putting interventions into four categories: multicomponent interventions, portfolio interventions, long causal chain interventions, and system-level interventions. These project types are intended as 'ideal types'. They are not meant to include all the characteristics of specific interventions, and real interventions may display elements of several project types. The project types were defined to clarify ideas and to attempt a pairing of project characteristics to appropriate evaluation methods. We acknowledge that our typology could be expanded or refined, but we believe our approach has been fruitful in terms of indicating what methods are useful to evaluate projects with given characteristics.

A limitation of our typology is that it does not provide a way to measure the degree of complexity of an intervention. We have argued that system-level interventions and portfolio interventions are harder to evaluate and that methods are available to evaluate interventions with long causal chains. However, this typology does not provide guidance to policymakers, managers, or researchers on how to decide whether an intervention is sufficiently complex to require the use of more methodologically sophisticated methods. Bamberger *et al.* (2016) have proposed a checklist that rates the level of complexity of an intervention along five key dimensions, which many will find useful to explore the complexity of interventions, to decide whether additional resources are needed, and to determine how the evaluation of different interventions should be prioritised.

As it is restricted to methods of causal inference, our review does not discuss mixed-method approaches. It seems obvious that interventions consisting of multiple components, operating at different levels, and implemented by different stakeholders should be analysed using a multiplicity of quantitative and qualitative methods. It would also be desirable for such methods to be integrated rather than employed in parallel, and for the conclusions to be triangulated rather than independently reported. Complex interventions are also more likely to involve various groups with different values and goals, and mixed-method approaches are best positioned to account for this diversity. We acknowledge that mixed-method approaches, their characteristics, and use should be the object of a separate review.

Finally, our review did not cover methods that we did not find used or mentioned in the literature. This has led to the omission of machine learning methods that have recently come to prominence. Machine learning methods are today applied to address complex scientific problems, and they could certainly support the evaluation of complex interventions. Some machine learning methods appear well positioned to support our understanding of complex interventions. For example, 'general additive models' are

designed precisely to model complex relationship between factors and their interactions, and are currently underutilised. Other methods, like tree-based methods, are currently used and explored, particularly in economics, and have found applications in understanding context heterogeneity of sensitive interventions. It is likely that researchers will draw increasingly from statistical innovations in these areas and apply them to development and evaluation problems. However, machine learning methods are inherently predictive and their use in the evaluation literature will require their integration with methods of causal inference, which implies this is an area of future methodological development.

7. Conclusions

The main goal of our paper was to identify methods to evaluate complex interventions. We started by building a typology of complex interventions and by pointing to the main challenges these interventions pose to evaluation. We then reviewed the (mostly) quantitative literature addressing causal inference in the evaluation of development interventions, and we identified promising methods that could be employed to evaluate specific project types. Here, we summarise the results of our review and provide some recommendations on which evaluation method should be used for each project type. These recommendations are summarised in Table 9.

Table 9: Recommended methods to evaluating complex interventions

Complex intervention	Recommended approaches
Multicomponent interventions	• Factorial designs • Adaptive trials • Agent-based modelling
Portfolio interventions	• Synthetic control • QCA
Long causal chain interventions	• Realist RCTs • Mechanism experiments • QCA
System-level intervention	• Agent-based modelling • System dynamics

We start with multicomponent interventions. Multicomponent interventions are projects consisting of various interacting components. The components are often evaluated separately or as a package, and evaluations fail to assess the impact of interactions between components—the 'synergistic' effects. We identified two methods from the experimental tradition that can support the evaluation of these projects: factorial designs and adaptive trials. These two methods are particularly well suited to evaluate interventions with activities that can be implemented in many different ways, and when there is much uncertainty about which particular activity or combination of activities is most effective. Both methods are currently heavily underutilised in international development, and we could find only three empirical applications.

Importantly, adaptive trials and factorial designs are not designed to evaluate the effectiveness of the interventions. They are better understood as components of a scientific process of search and optimisation of effective practices. Their increased use would shift the current focus of evaluators from the effectiveness stage of an intervention to the design stage. Their primary goal is to help researchers and

managers in designing and optimising interventions. They are designed to support the evaluation of interventions where there is limited knowledge of the operating mechanisms or where there are many different potential implementation strategies. Given that ignorance about operating mechanisms is common, as is having a variety of implementation options it is perhaps surprising that these methods are not used more frequently and that the standard RCT has become the preferred method of choice in much of the impact evaluation literature.

We include agent-based modelling as another potential method for the prospective evaluation of multicomponent interventions. Agent-based modelling performs simulations of interventions rather than assessing effectiveness, but we found a few cases in our review that addressed the estimation of interaction effects between project components. The ability of simulating the impact of interactions between components across different sectors is not limited to agent-based modelling and belongs to modelling approaches more generally, of which agent-based modelling is one example. Structural modelling, in particular, as practised in economics, is similarly well positioned to conduct this type of evaluation.

The second type of complex intervention considered were portfolio interventions. These are interventions consisting of multiple components, often implemented across multiple sectors, that have the additional complication of being implemented in a whole country or across several countries. The scale of the intervention represents an additional challenge to evaluation, which translates into the inability of identifying a valid control group. These projects are often evaluated using qualitative methods or using quantitative methods that do not employing a counterfactual such as before–after analyses.

We have recommended the use of synthetic control methods for the evaluation of portfolio interventions. Synthetic control uses time series data but, unlike interrupted time series, they build a valid counterfactual. They address the thorny issue of assessing impacts of interventions occurring at regional or country level and spanning several years. Synthetic control is the only truly relatively new method covered in our review and, as such, it is constantly being developed and improved.

Another method for the evaluation of portfolio interventions is QCA. QCA, like synthetic control, originated in comparative case analysis. Like synthetic control, it requires a deep knowledge of the cases analysed and uses a small number of observations. Unlike synthetic control, however, it does not rely on time series data. lt pays particular attention to contextual factors and combinations of characteristics that are associated with the success—and lack of success—of interventions. While it is weaker than

synthetic control in terms of causal inference, it has the advantage of providing some understanding of the causes of effects.

In our review, we have not discussed methods for the evaluation of long causal chain interventions. These are interventions that try to change multiple behaviours through long causal chains, and which are highly sensitive to context characteristics. The reason for not covering this type of project is that many available methods are currently used to evaluate this type of projects effectively. Even simple RCTs, if framed in the form of 'mechanism experiments', can be employed in the evaluation of this type of intervention. More often, RCTs are carried out alongside process evaluations in 'pragmatic' or 'realist' trials, which not only assess the effectiveness of the interventions but also inform how they work. What is true of RCTs also applies to quasi-experimental methods used in conjunction with qualitative methods in mixed-method evaluations.

We included QCA in the recommended list of methods evaluating projects with long causal chains. The reason is that, in our review, we found QCA was often used to explore the interaction of project activities with the characteristic of the context in which the projects were implemented. One of the main challenges of interventions with long causal chains is the high sensitivity of the activities and outcomes to contextual characteristics of implementation. QCA appears able to shed some light on project-context interactions, and could be effectively employed together with other evaluation methods addressing causal inference.

System-level interventions are the complex interventions that are most difficult to evaluate. System-level interventions aim at changing the way a 'whole system' works. Because of the large scale at which they operate, because of the many outcomes and sectors of interventions, and because of the possibility of producing emergent outcomes, they cannot be evaluated with standard quantitative methods. We reviewed two modelling approaches developed to understand the operation of complex systems: agent-based modelling and system dynamics. Agent-based modelling is well positioned to evaluate system-level interventions. We are more hesitant about system dynamics, which seems more indicated to understanding the context of intervention and to supporting project design.

Both approaches, and system dynamic in particular, have some limitations in the way their models are validated with the data and in how robustness analysis is conducted. In general, the results of modelling methods are credible to the extent the models can predict the observed data with reasonable accuracy, and to the extent they are not too sensitive to small changes in specification and parameter. However, both agent-based models and system dynamics are not strong in either validation or robustness. The

development of methods of validation and robustness analysis of these methods would greatly contribute to their credibility and wider use.

We conclude with a final note on the risk of 'complexification' in the evaluation of development interventions. It is sometimes claimed that all development interventions are complex, and that evaluations should take a complexity perspective. We have argued in the paper that this is only true to an extent. Our review of factorial studies, for example, showed that synergistic interactions, the simplest form of project complexity, are relatively rare and that project effects appear surprisingly linear. It is possible that the 'linear' results of factorial trials are driven by an excessive focus on short-term impacts and on simple interventions, but these results should introduce a note of caution. Linear effects are possibly more common than is normally thought, at least within the time horizon and the types of project that are of interest to decision makers. More importantly, the complexity of the context should be established and verified rather than simply being assumed. The purported complexity of interventions should not become a pretext for not rigorously evaluating large-scale programmes.

We have indicated a series of available evaluation methods that can be used in some circumstances. Factorial designs and adaptive trials are experimental methods that could be effectively employed in the evaluation of complex multicomponent interventions at their design stage. Quasi-experimental designs, such as synthetic control and its developments, and QCA can be used in the evaluation of large-scale portfolio interventions. No issues of cost, feasibility, or rigour prevent the use of these methods.

Modelling methods, such as agent-based modelling and system dynamics, can contribute to the evaluation of system-level intervention. The system approaches currently in use in the evaluation literature are useful to the understanding of complex interventions and to their design, but are not able to assess their impact. Modelling approaches can support impact assessment if they are employed alongside counterfactual designs, and if adequate methods of validation and robustness analysis are conducted. The use of system-level approaches in impact evaluation requires that these methods are further developed and refined in light of the causal inference literature. This integration between system-level approaches and causal inference is potentially an area for future methodological development, and one to which researchers have so far given very limited attention.

Another area of potential development is the integration between machine learning and methods of causal inference. Although not covered in our review, machine learning methods are designed to deal with large datasets and with complex relationship

between many variables. These methods, however, are entirely predictive, and not designed to provide causal explanations. Developments in computing and interactions between researchers of different disciplines might make the integration between machine learning and causal inference methods another area of future development.

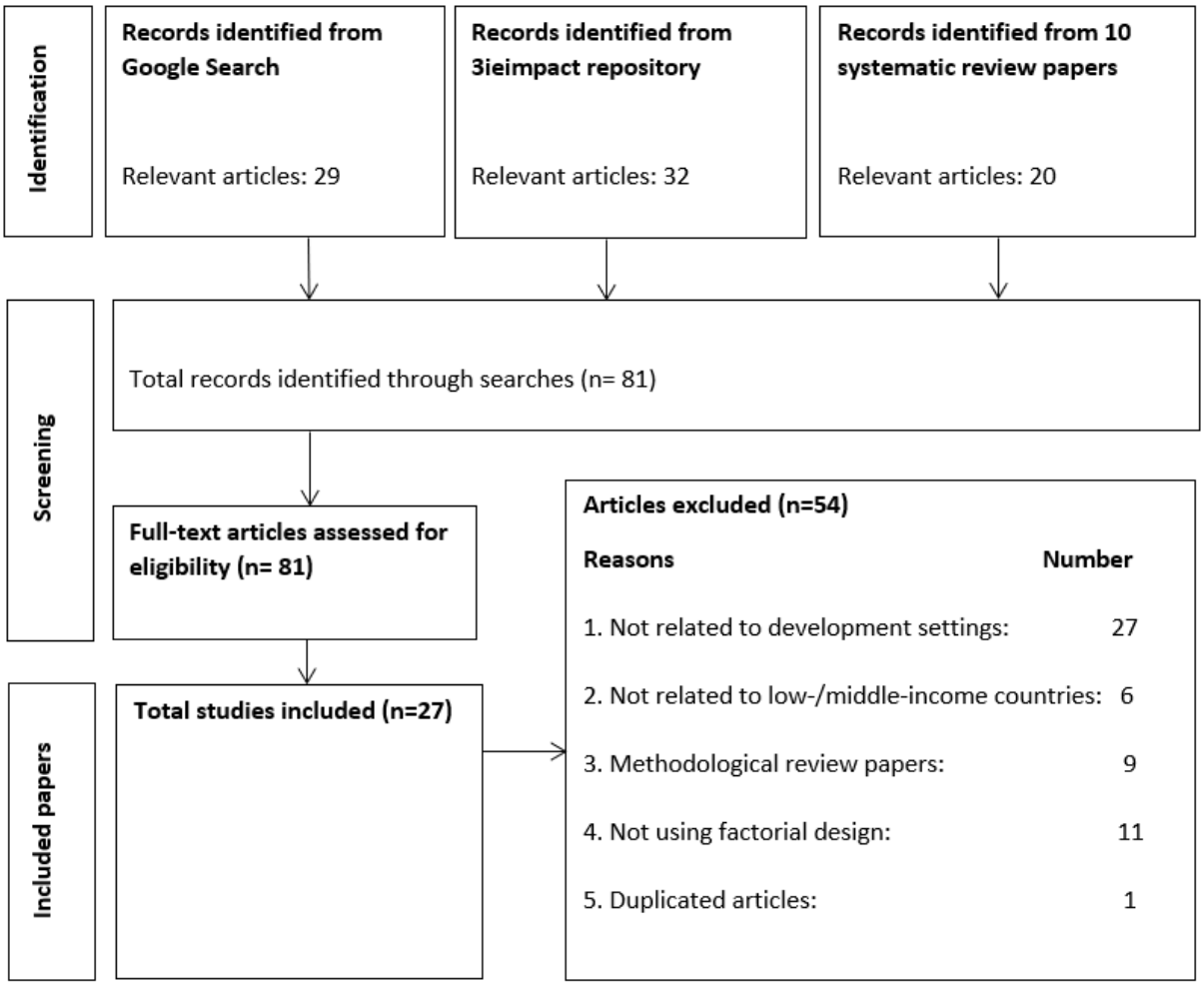
Finally, our review has not explored evaluation issues related to the complexity of the environment in which the interventions are implemented. As discussed in the review, contextual complexity generates problems of external validity and further challenges in the evaluation of interventions that change over time by adapting to changing environmental circumstances. Methods to address external validity are a lively area of methodological research, but much less research is likely to be available regarding the evaluation of interventions implemented in complex environments. This could be the topic of another review.

8. Appendix

Factorial designs

We identified 10 systematic review articles and reports and 29 relevant articles and reports through a Google search. We found 32 articles and evaluation reports in the 3ie repository. The review of the systematic reviews led to another 20 potential articles for inclusion. Thus, a total of 81 articles and reports were fully assessed by reading the abstracts and text. A total of 27 articles that fulfilled the inclusion criteria were ultimately selected for review. Details of the selection process are reported in the flow diagram below.

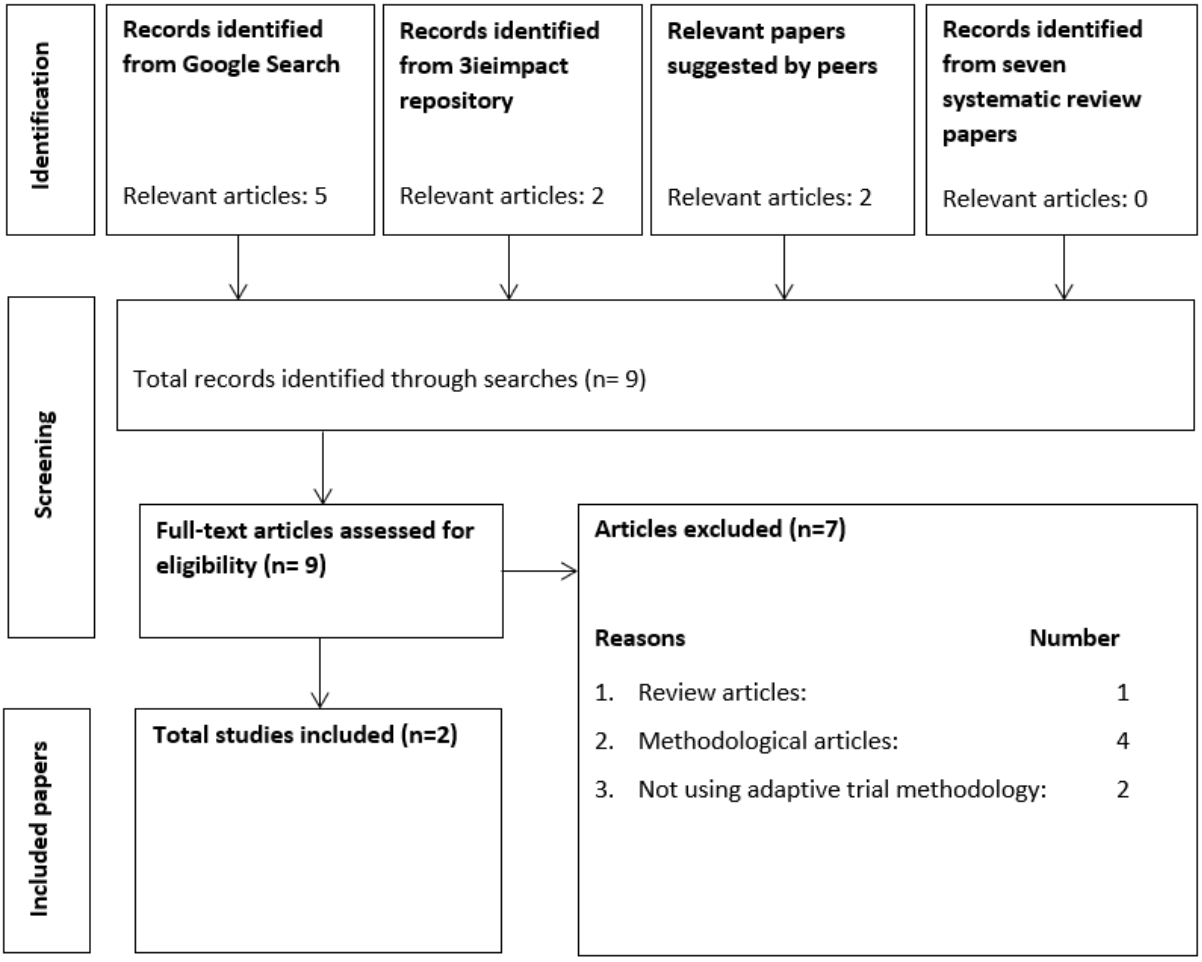
Figure A1 Flow diagram of literature search: factorial designs



Adaptive trials

We identified seven systematic reviews and five articles through the Google search. We identified two further articles in the 3ie repository, and two further articles were referred by expert colleagues. The reading of systematic reviews did not provide any result, because most articles included in the reviews consisted of drug efficacy trials or biomedical interventions. Therefore, a total of nine studies were selected for full paper review. The application of selection criteria led to the exclusion of seven papers, and only two paper were eventually selected for a full review. The details of the search strategy and results are provided in the figure below.

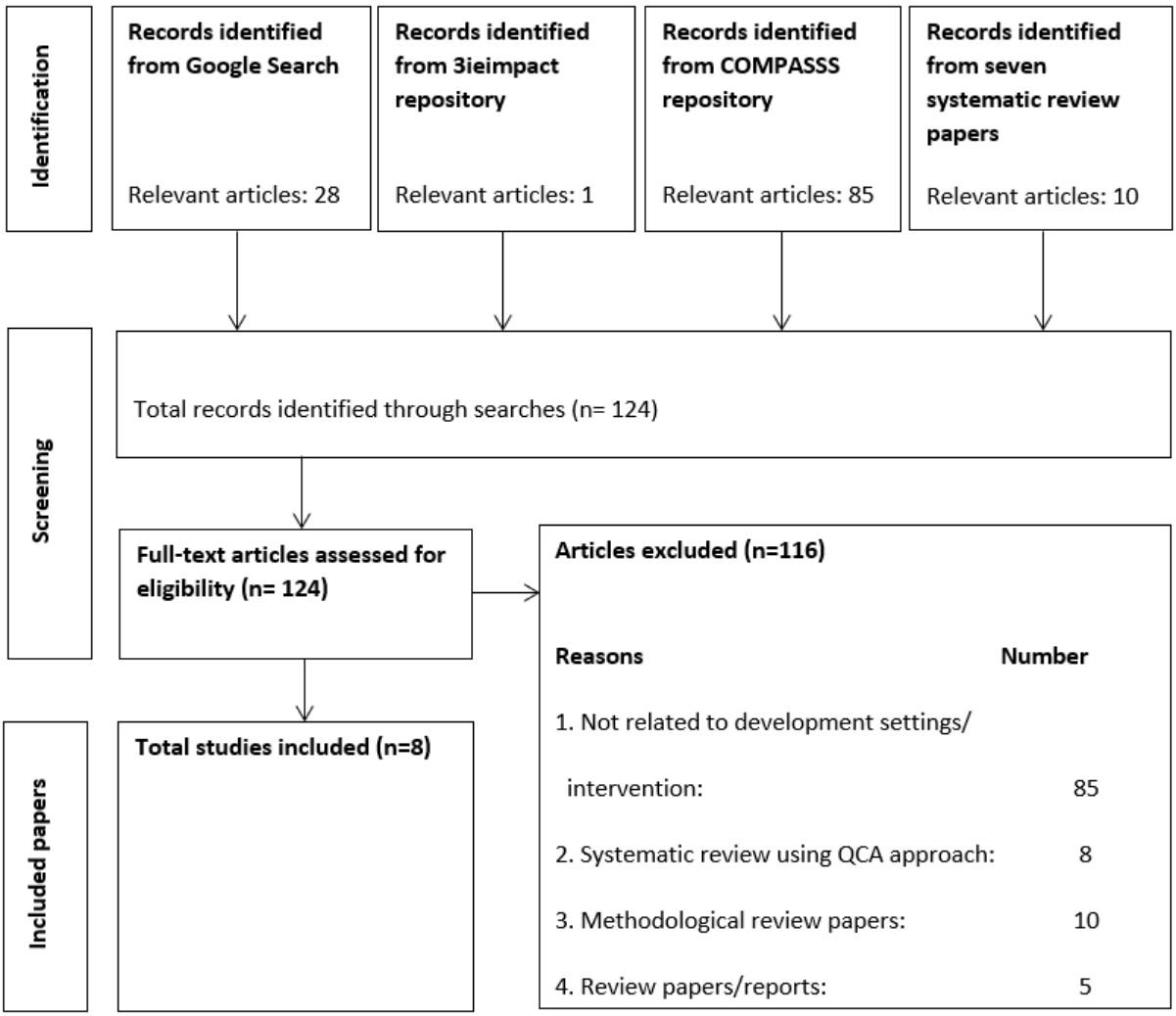
Figure A2 Flow diagram of literature search: adaptive trials



QCA

We identified seven systematic reviews and 28 articles through our Google search. In addition, we identified 85 articles and reports in the COMPASSS repository, and one article from the 3ie repository. After examining the seven systematic reviews, we identified 10 additional articles. We screened text and the abstracts of 124 articles and selected eight for the final review. The details on the search and results are provided in the figure below.

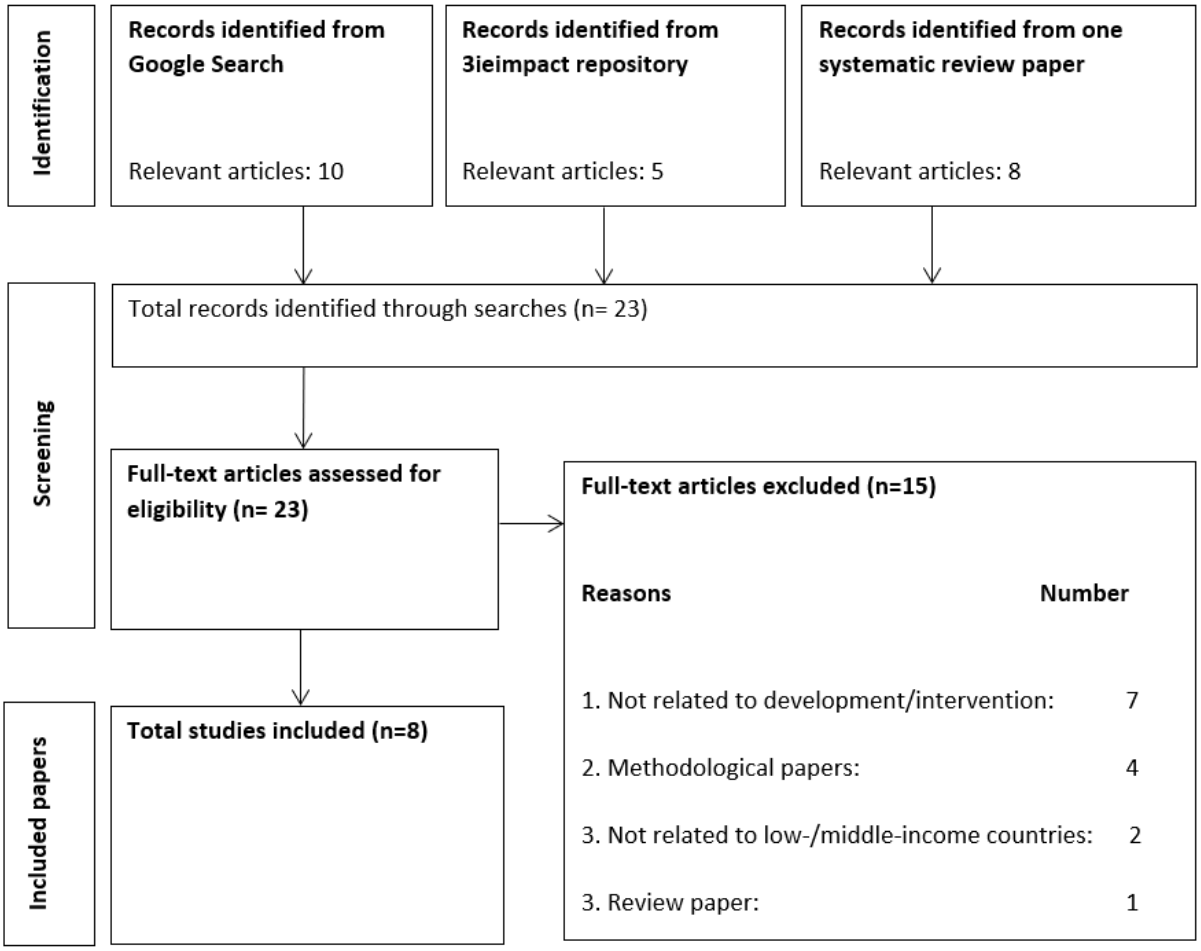
Figure A3 Flow diagram of literature search: QCA



Synthetic control

We identified one systematic review and 10 articles through Google search, and five more articles were found in the 3ie repository. We extracted eight additional papers from the systematic reviews. In total, 23 articles were fully assessed for inclusion. After a full-text review, 15 articles were excluded and only eight were included. Details of the search and results are reported in the figure below.

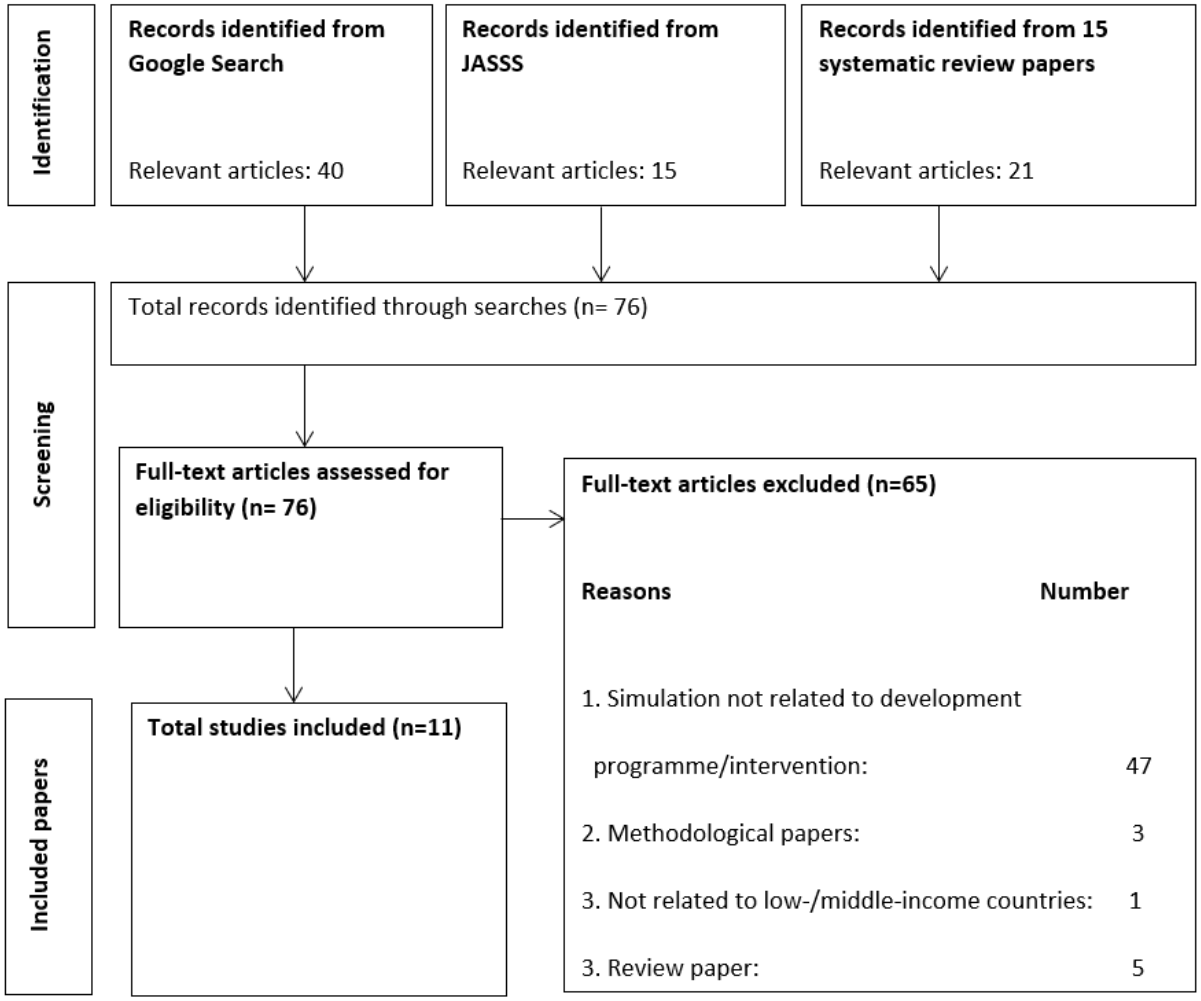
Figure A4 Flow diagram of literature search: synthetic control methods



Agent-based modelling

We found 15 systematic reviews through our Google search. We found 40 primary studies through our Google search and a further 15 articles in JASSS. After reviewing the systematic review articles, we identified a further 21 potential articles. In total, 76 articles were fully reviewed and 11 were retained. The details of the search and its results are reported in the figure below.

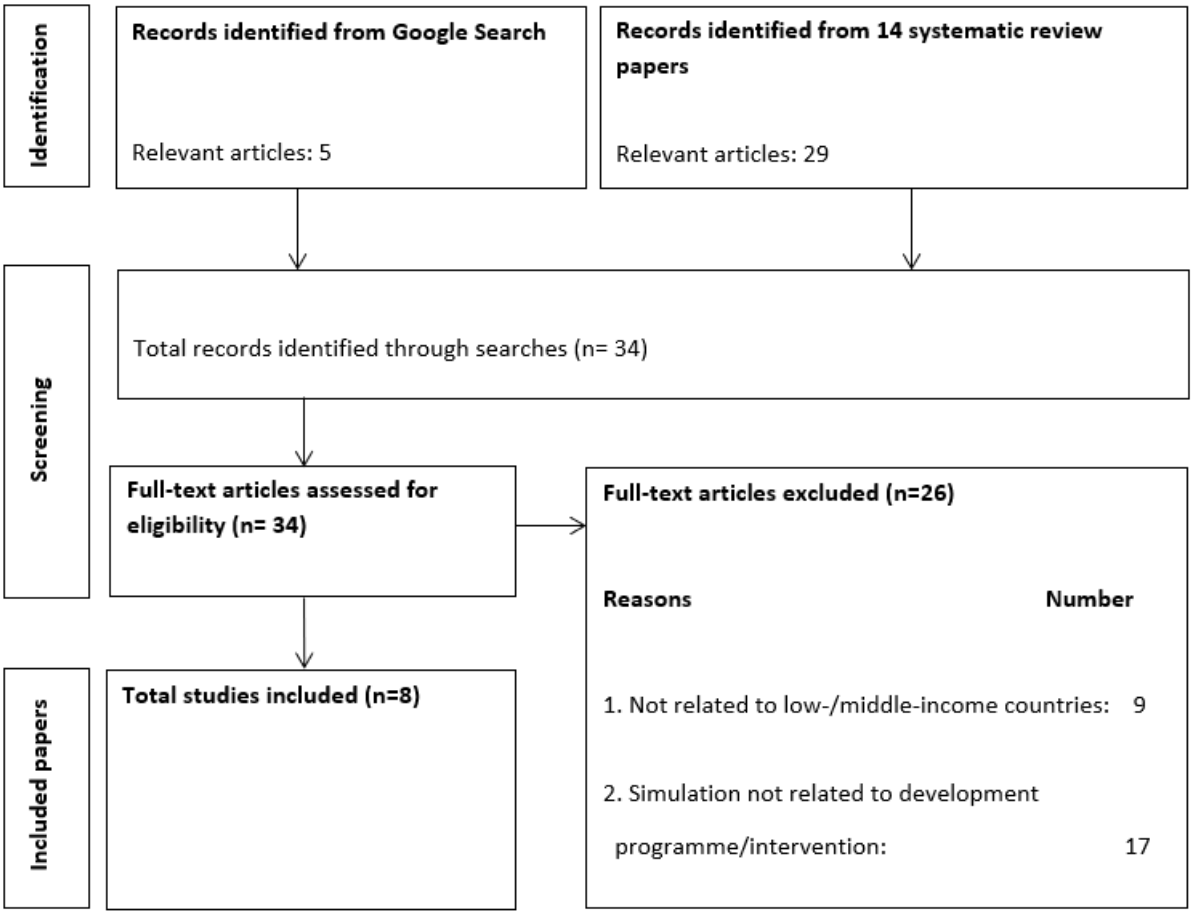
Figure A5 Flow diagram of literature search agent-based modelling



System dynamics

We found 14 systematic reviews and five primary studies through our Google search. After reviewing the systematic reviews, 29 primary studies were selected for further assessment. In total, 34 papers were fully reviewed and eight were selected after the application of eligibility criteria. The details of the search and the results are reported in the figure below.

Figure A6 Flow diagram of literature search: system dynamic modelling



9. References

Abadie, A. (2020). Using synthetic controls: feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*.

Abadie, A., Diamond, A. and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), pp. 493–505. doi: 10.1198/jasa.2009.ap08746

Abadie, A., Diamond, A. and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2), pp. 495–510. doi: 10.1111/ajps.12116

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: a case study of the Basque Country. *American Economic Review*, 93(1), pp. 113–132. doi: 10.1257/000282803321455188

Al-Khatib, I.A., Eleyan, D. and Garfield, J. (2016). A system dynamics approach for hospital waste management in a city in a developing country: the case of Nablus, Palestine. *Environmental Monitoring and Assessment*, 188(9), p. 503. doi: 10.1007/s10661-016-5487-9

Alavuotunki, K. (2015). *General Budget Support, Health Expenditures, and Neonatal Mortality Rate: A Synthetic Control Approach*. The United Nations University World Institute for Development Economics Research (UNU-WIDER) (WIDER Working Paper), Helsinki. doi: 10.35188/UNU-WIDER/2015/997-8

Alonge, O. *et al.* (2017). Improving health systems performance in low- and middle-income countries: a system dynamics model of the pay-for-performance initiative in Afghanistan. *Health Policy and Planning*, 32(10), pp. 1417–1426. doi: 10.1093/heapol/czx122

An, L. *et al.* (2020). Cascading impacts of payments for ecosystem services in complex human-environment systems. *Journal of Artificial Societies and Social Simulation*, 23(1). doi: 10.18564/jasss.4196

An, Y., Garvin, M.J. and Hall, R.P. (2017). Pathways to better project delivery: the link between capacity factors and urban infrastructure projects in India. *World Development*, 94, pp. 393–405. doi: 10.1016/j.worlddev.2017.02.004

Andersen, L. *et al.* (2014). *Modelling Land Use, Deforestation, and Policy Analysis: A Hybrid Optimization-ABM Heterogeneous Agent Model with Application to the Bolivian Amazon*. LSE: Centre for Climate Change Economics and Policy, London [online]. Available from: www.lse.ac.uk/GranthamInstitute/wp-content/uploads/2014/09/Working-Paper-164-Andersen-et-al-2015.pdf (Accessed 6 October 2021).

Arkhangelsky, D. *et al.* (2018). Synthetic difference in differences. *arXiv*.

Athey, S. and Imbens, G.W. (2017). The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), pp. 3–32. doi: 10.1257/jep.31.2.3

Baker, T.B. *et al.* (2017). Implementing clinical research using factorial designs: a primer. *Behavior Therapy*, 48(4), pp. 567–580. doi: 10.1016/j.beth.2016.12.005

Bamberger, M., Vaessen, J. and Raimondo, E. (2016). *Dealing With Complexity in Development Evaluation: A Practical Approach*. SAGE Publications, Thousand Oaks, CA. doi: 10.4135/9781483399935

Bandiera, O. *et al.* (2017). Labor markets and poverty in village economies. *The Quarterly Journal of Economics*, 132(2), pp. 811–870. doi: 10.1093/qje/qjx003

Banerjee, A. *et al.* (2015). A multifaceted program causes lasting progress for the very poor: evidence from six countries. *Science*, 348(6236), pp. 1260799–1260799. doi: 10.1126/science.1260799

Barnes, M., Matka, E. and Sullivan, H. (2003). Evidence, understanding and complexity. *Evaluation*, 9(3), pp. 265–284. doi: 10.1177/13563890030093003

Barofsky, J., Anekwe, T.D. and Chase, C. (2015). Malaria eradication and economic outcomes in sub-Saharan Africa: evidence from Uganda. *Journal of Health Economics*, 44, pp. 118–36. doi: 10.1016/j.jhealeco.2015.08.002

Beach, D. (2017). Process-tracing methods in social science. In: *Oxford Research Encyclopedia of Politics*. Oxford University Press, Oxford. doi: 10.1093/acrefore/9780190228637.013.176

Belem, M., Bazile, D. and Coulibaly, H. (2018). Simulating the impacts of climate variability and change on crop varietal diversity in Mali (West-Africa) using Agent-Based Modeling approach. *Journal of Artificial Societies and Social Simulation*, 21(2). doi: 10.18564/jasss.3690

Van Belle, S. *et al.* (2016). Can 'realist' randomised controlled trials be genuinely realist? *Trials*, 17(1), p. 313. doi: 10.1186/s13063-016-1407-0

Bifulco, R., Rubenstein, R. and Sohn, H. (2017). Using synthetic controls to evaluate the effect of unique interventions: the case of Say Yes to Education. *Evaluation Review*. doi: 10.1177/0193841X17742233

Birdthistle, I. et al. (2018) 'Evaluating the impact of the DREAMS partnership to reduce HIV incidence among adolescent girls and young women in four settings: a study protocol', BMC Public Health, 18(1), p. 912. doi: 10.1186/s12889-018-5789-7.

Blattman, C., Jamison, J.C. and Sheridan, M. (2017). Reducing crime and violence: experimental evidence from Cognitive Behavioral Therapy in Liberia. *The American Economic Review*, 107(4), pp. 1165–1206. doi: 10.1257/aer.20150503

Bonell, C. *et al.* (2016). Realist trials and the testing of context-mechanism-outcome configurations: a response to Van Belle *et al*. *Trials*, 17(1), p. 478. doi: 10.1186/s13063-016-1613-9

Borgomeo, E., Hall, J.W. and Salehin, M. (2018). Avoiding the water-poverty trap: insights from a conceptual human-water dynamical model for coastal Bangladesh. *International Journal of Water Resources Development*, 34(6), pp. 900–922. doi: 10.1080/07900627.2017.1331842

Bothwell, L.E. *et al.* (2018). Adaptive design clinical trials: a review of the literature and ClinicalTrials.gov. *BMJ Open*, 8(2), p. e018320. doi: 10.1136/bmjopen-2017-018320

Bouttell, J. *et al.* (2018). Synthetic control methodology as a tool for evaluating population-level health interventions. *Journal of epidemiology and community health*, 72(8), pp. 673–678. doi: 10.1136/jech-2017-210106

Box, G.E.P., Hunter, J.S. and Hunter, W.G. (2005). *Statistics of Experimenters (2nd ed.)* John Wiley & sons.

Breuer, E. *et al.* (2018). Using qualitative comparative analysis and theory of change to unravel the effects of a mental health intervention on service utilisation in Nepal. *BMJ Global Health*, 3(6), p. e001023. doi: 10.1136/bmjgh-2018-001023

Brodersen, K.H. *et al.* (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1). doi: 10.1214/14-AOAS788

Brück, T. *et al.* (2019). Assets for alimentation? The nutritional impact of assets-based programming in Niger. *The Journal of Development Studies*, 55(sup1), pp. 55–74. doi: 10.1080/00220388.2019.1687876

Buffardi, A.L. and Hearn, S. (2015). *Multi-Project Programmes: Functions, Forms, and Implications for Evaluation and Learning*. London.

Burke, H.M., Chen, M. and Brown, A.N. (2018). Applying factorial designs to disentangle the effects of integrated development. *IDS Bulletin*, 49(4). doi: 10.19088/1968-2018.165

Campbell, M. *et al.* (2000). Framework for design and evaluation of complex interventions to improve health. *British Medical Journal*. doi: 10.1136/bmj.321.7262.694

Caria, S. *et al.* (2020). *An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan*. CESifo Working Paper.

Cartwright, N. *et al.* (2020). *Making Predictions of Programme Success More Reliable*. CEDIL Methods Working Paper. CEDIL, Oxford.

Cartwright, N. and Hardie, J. (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford University Press, Oxford.

Cassidy, R. *et al.* (2019a). Mathematical modelling for health systems research: a systematic review of system dynamics and Agent-Based Models. *BMC health services research*, 19(1), p. 845. doi: 10.1186/s12913-019-4627-7

Cassidy, R. *et al.* (2019b). Mathematical modelling for health systems research: a systematic review of system dynamics and Agent-Based Models. *BMC Health Services Research*, 19(1), p. 845. doi: 10.1186/s12913-019-4627-7

Chatterley, C. *et al.* (2014). A qualitative comparative analysis of well-managed school sanitation in Bangladesh. *BMC Public Health*, 14(1), p. 6. doi: 10.1186/1471-2458-14-6

Chelwa, G., van Walbeek, C. and Blecher, E. (2017). Evaluating South Africa's tobacco control policy using a synthetic control method. *Tobacco Control*, 26(5), pp. 509–517. doi: 10.1136/tobaccocontrol-2016-053011

Collins, L.M. *et al.* (2014). Factorial Experiments. *American Journal of Preventive Medicine*, 47(4), pp. 498–504. doi: 10.1016/j.amepre.2014.06.021

Craig, P. *et al.* (2019). *Developing and Evaluating Complex Interventions*. Medical Research Council, London.

Davies, R. (2018). Representing theories of change: technical challenges with evaluation consequences. *Journal of Development Effectiveness*, 10(4), pp. 438–461. doi: 10.1080/19439342.2018.1526202

Davis, A., Javernick-Will, A. and Cook, S.M. (2019). The use of qualitative comparative analysis to identify pathways to successful and failed sanitation systems. *Science of The Total Environment*, 663, pp. 507–517. doi: 10.1016/j.scitotenv.2019.01.291

DFID (2016) *Business Case: Jordan Compact Economic Opportunities Programme*. URL:https://iati.fcdo.gov.uk/iati_documents/5519561.odt Accessed 2021/10/07

Ding, Z. *et al.* (2018). System dynamics versus Agent-Based Modeling: a review of complexity simulation in construction waste management. *Sustainability*, 10(7), p. 2484. doi: 10.3390/su10072484

Du, L. *et al.* (2015). Leveraging agriculture for nutrition impact through the Feed the Future Initiative. In: *Advances in Food and Nutrition Research*, pp. 1–46. doi: 10.1016/bs.afnr.2014.11.001

Duflo, E., Glennerster, R. and Kremer, M. (2007). Using randomization in development economics research: a toolkit. In: *Handbook of Development Economics*, Chapter 61 (pp. 3895–3962). doi: 10.1016/S1573-4471(07)04061-2

Dunning, T. *et al.* (2019). *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge University Press, Cambridge.

Egger, M. *et al.* (2000). Promotion of condom use in a high-risk setting in Nicaragua: a randomised controlled trial. *The Lancet*, 355(9221), pp. 2101–2105. doi: 10.1016/S0140-6736(00)02376-X

Elbers, C. and Gunning, J.W. (2014). Evaluation of development programs: randomized controlled trials or regressions? *The World Bank Economic Review*, 28(3), pp. 432–445. doi: 10.1093/wber/lht025

Elbers, C., Gunning, J.W. and de Hoop, K. (2009). Assessing sector-wide programs with statistical impact evaluation: a methodological proposal. *World Development*, 37(2), pp. 513–520. doi: 10.1016/j.worlddev.2008.01.002

Elliot, K. and Dunning, C. (2016). *Assessing the US Feed the Future Initiative: A New Approach to Food Security?* CGD Policy Paper 075.

Feola, G., Gallati, J.A. and Binder, C.R. (2012). Exploring behavioural change through an agent-oriented system dynamics model: the use of personal protective equipment among pesticide applicators in Colombia. *System Dynamics Review*, 28(1), pp. 69–93. doi: 10.1002/sdr.469

Forrester, J.W. (1962). Industrial dynamics. *Journal of the American Statistical Association*, 57(298), p. 525. doi: 10.2307/2281677

Franco, L. et al. (2019) Big Bet On Nigeria Synthesis Report. Rockville, MD: EnCompass LLC. Available at: https://www.macfound.org/media/files/2018_synthesisreport_2019.05.30_public_final.pdf.

Fredericks, K.A., Deegan, M. and Carman, J.G. (2008). Using system dynamics as an evaluation tool. *American Journal of Evaluation*, 29(3), pp. 251–267. doi: 10.1177/1098214008319446

García, P.J. *et al.* (2012). Prevention of sexually transmitted infections in urban communities (Peru PREVEN): a multicomponent community-randomised controlled trial. *The Lancet*, 379(9821), pp. 1120–1128. doi: 10.1016/S0140-6736(11)61846-1

Gates, E.F. (2018). Toward valuing with critical systems heuristics. *American Journal of Evaluation*. doi: 10.1177/1098214017703703

Gharakhanlou, N.M., Hooshangi, N. and Helbich, M. (2020). A spatial Agent-Based Model

to assess the spread of malaria in relation to anti-malaria interventions in southeast Iran. *ISPRS International Journal of Geo-Information*, 9(9), p. 549. doi: 10.3390/ijgi9090549

Glouberman, S. and Zimmerman, B. (2002). *Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like?* Change.

Greenhalgh, T. and Papoutsi, C. (2018). Studying complexity in health services research: desperately seeking an overdue paradigm shift. *BMC Medicine*, 16(1), p. 95. doi: 10.1186/s12916-018-1089-4

Gutierrez, I., Weinberger, G. and Engberg, J. (2016). *Improving Teaching Effectiveness: Impact on Student Outcomes: The Intensive Partnerships for Effective Teaching Through 2013–2014*. RAND Corporation. doi: 10.7249/RR1295.3

Hassmiller Lich, K. *et al.* (2010). Using system dynamics tools to gain insight into intervention options related to the interaction between tobacco and tuberculosis. *Global Health Promotion*, 17(1 Suppl), pp. 7–20. doi: 10.1177/1757975909358255

HM Treasury (2020). *Magenta Book 2020 Supplementary Guide: Handling Complexity in Policy Evaluation*. UK Government, London.

Hummelbrunner, R. (2011). Systems thinking and evaluation. *Evaluation*, 17(4), pp. 395–403. doi: 10.1177/1356389011421935

Imbens, G.W. and Rubin, D.B. (2015). *Causal Inference: For Statistics, Social, and Biomedical Sciences an Introduction*. Cambridge University Press, Cambridge. doi: 10.1017/CBO9781139025751

IOM (2013) Evaluation of PEPFAR. Washington DC: The National Academies Press. Available at: http://www.nap.edu/catalog.php?record_id=18256.

Jafar, T.H. *et al.* (2009). Community-based interventions to promote blood pressure control in a developing country: a cluster randomized trial. *Annals of Internal Medicine*, 151(9), pp. 593–601. doi: 10.7326/0003-4819-151-9-200911030-00004

Jamal, F. *et al.* (2015). The three stages of building and testing mid-level theories in a realist RCT: a theoretical and methodological case-example. *Trials*, 16(1), pp. 1–10. doi: 10.1186/S13063-015-0980-Y

Jones, M. *et al.* (2016). OPALesence: epistemological pluralism in the evaluation of a systems-wide childhood obesity prevention program. *Evaluation*, 22(1), pp. 29–48. doi: 10.1177/1356389015623142

Juneja, S. *et al.* (2013). Impact of an HIV prevention intervention on condom use among long distance truckers in India. *AIDS and Behavior*, 17(3), pp. 1040–1051. doi: 10.1007/s10461-012-0314-y

Kasy, M. and Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1), pp. 113–132. doi: 10.3982/ECTA17527

Koleros, A. *et al.* (2016). Methodological considerations in evaluating long-term systems change. *American Journal of Evaluation*, 37(3), pp. 364–380. doi: 10.1177/1098214015615231

Koleros, A. *et al.* (2020). The actor-based change framework: a pragmatic approach to developing program theory for interventions in complex systems. *American Journal of Evaluation*, 41(1), pp. 34–53. doi: 10.1177/1098214018786462

Kremmydas, D., Athanasiadis, I.N. and Rozakis, S. (2018). A review of Agent Based

Modeling for agricultural policy evaluation. *Agricultural Systems*, 164, pp. 95–106. doi: 10.1016/j.agsy.2018.03.010

Lachman, J.M. *et al.* (2019). Preventing child mental health problems through parenting interventions in Southeastern Europe (RISE): protocol for a multi-country cluster randomized factorial study. *Contemporary Clinical Trials*, 86, p. 105855. doi: 10.1016/j.cct.2019.105855

Lam, W.F. and Ostrom, E. (2010). Analyzing the dynamic complexity of development interventions: lessons from an irrigation experiment in Nepal. *Policy Sciences*, 43(1), pp. 1–25. doi: 10.1007/s11077-009-9082-6

Lemoine, P.D. *et al.* (2016). Using Agent Based Modeling to assess the effect of increased Bus Rapid Transit system infrastructure on walking for transportation. *Preventive Medicine*, 88, pp. 39–45. doi: 10.1016/j.ypmed.2016.03.015

Lepine, A., Lagarde, M. and Le Nestour, A. (2014). *Free Primary Care in Zambia: An Impact Evaluation Using a Pooled Synthetic Control Method*, *SSRN Electronic Journal*. HEDG Working Paper, University of York. doi: 10.2139/ssrn.2520345

Li, M. *et al.* (2019). System dynamics modeling of public health services provided by China CDC to control infectious and endemic diseases in China. *Infection and Drug Resistance*, 12, pp. 613–625. doi: 10.2147/IDR.S185177

Li, X., Sudarsanam, N. and Frey, D.D. (2006). Regularities in data from factorial experiments. *Complexity*, 11(5), pp. 32–45. doi: 10.1002/cplx.20123

Lieberson, S. and Ragin, C.C. (2001). Fuzzy-set social science. *Contemporary Sociology*, 30(4), p. 331. doi: 10.2307/3089736

Lipton, M. (2010). *New Seeds and Poor People*. Routledge. doi: 10.4324/9780203840436

Ludwig, J., Kling, J.R. and Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives*. doi: 10.1257/jep.25.3.17

Mackenzie, J. and Hearn, S. (2016) Impact evaluation for portfolio programmes on policy influence. London: A Methods Lab publication. Overseas Development Institute. Available at: https://cdn.odi.org/media/documents/10463.pdf.

Mackie, J.L. (1965). Causes and conditions. *American Philosophical Quarterly*, 2(4), pp. 245–264.

Marshall, B.D.L. and Galea, S. (2015). Formalizing the role of Agent-Based Modeling in causal inference and epidemiology. *American Journal of Epidemiology*, 181(2), pp. 92–9. doi: 10.1093/aje/kwu274

Masset, E., García-Hombrados, J. and Acharya, A. (2020). Aiming high and falling low: the SADA-Northern Ghana Millennium Village Project. *Journal of Development Economics*, 143, p. 102427. doi: 10.1016/j.jdeveco.2019.102427

Mayne, J. (2001). Addressing attribution through contribution analysis: using performance measures sensibly. *The Canadian Journal of Program Evaluation*.

Mbiti, I. *et al.* (2019). Inputs, incentives, and complementarities in education: experimental evidence from Tanzania. *The Quarterly Journal of Economics*, 134(3), pp. 1627–1673. doi: 10.1093/qje/qjz010

McAlpine, A. *et al.* (2020). Agent-Based Modeling for migration and modern slavery research: a systematic review. *Journal of Computational Social Science*. doi:

10.1007/s42001-020-00076-7

McGill, E. *et al.* (2021). Evaluation of public health interventions from a complex systems perspective: a research methods review. *Social Science & Medicine*, 272, p. 113697. doi: 10.1016/J.SOCSCIMED.2021.113697

Mdege, N. D. et al. (2014) 'The 2 × 2 cluster randomized controlled factorial trial design is mainly used for efficiency and to explore intervention interactions: a systematic review.', Journal of clinical epidemiology, 67(10), pp. 1083–92. doi: 10.1016/j.jclinepi.2014.06.004.

van Mierlo, B., Arkesteijn, M. and Leeuwis, C. (2010). Enhancing the reflexivity of system innovation projects with system analyses. *American Journal of Evaluation*, 31(2), pp. 143–161. doi: 10.1177/1098214010366046

Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press, New York.

Moore, G.F. *et al.* (2015). Process evaluation of complex interventions: Medical Research Council guidance. *BMJ*, 350(mar19 6), pp. h1258–h1258. doi: 10.1136/bmj.h1258

Moore, G.F. *et al.* (2019). From complex social interventions to interventions in complex social systems: future directions and unresolved questions for intervention development and evaluation. *Evaluation*, 25(1), pp. 23–45. doi: 10.1177/1356389018803219

Moore, M., Parsons, B. and Jessup, P. (2019). How evaluators can use a complex systems lens to get 'untrapped' From limiting beliefs and assumptions. *American Journal of Evaluation*, 40(1), pp. 75–91. doi: 10.1177/1098214018756578

Mowles, C. (2014). Complex, but not quite complex enough: the turn to the complexity sciences in evaluation scholarship. *Evaluation*, 20(2), pp. 160–175. doi: 10.1177/1356389014527885

MRC (2000). *A Framework for Development and Evaluation of RCTs for Complex Interventions to Improve Health*. Medical Research Council, London.

Muralidharan, K., Romero, M. and Wüthrich, K. (2019). *Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments.* NBER Working Paper Series.

Nguyen, H.K. *et al.* (2019). Contract farming in the Mekong Delta's rice supply chain: insights from an Agent-Based Modeling study. *Journal of Artificial Societies and Social Simulation*, 22(3). doi: 10.18564/jasss.4008

Nianogo, R.A. and Arah, O.A. (2015). Agent-Based Modeling of noncommunicable diseases: a systematic review. *American Journal of Public Health*, 105(3), pp. e20-31. doi: 10.2105/AJPH.2014.302426

Nöldeke, B., Winter, E. and Grote, U. (2020). Seed selection strategies for information diffusion in social networks: an Agent-Based Model applied to rural Zambia. *Journal of Artificial Societies and Social Simulation*, 23(4). doi: 10.18564/jasss.4429

Nsubuga, P., Polhemus, M. and Gaur, D. (2016) External Evaluation of the USAID Malaria Vaccine Development Program. USAID Global Health Performance Cycle Improvement Project. Available at: https://www.usaid.gov/sites/default/files/documents/1864/MVDP-External-Evaluation-May-2016-508.pdf.

Oliver, S. *et al.* (2021). *Engaging Stakeholders with Evidence and Uncertainty: Developing a Toolkit.* CEDIL Meth. CEDIL, Oxford.

Opdyke, A., Javernick-Will, A. and Koschmann, M. (2019). Assessing the impact of

household participation on satisfaction and safe design in humanitarian shelter projects. *Disasters*, 43(4), pp. 926–953. doi: 10.1111/disa.12405

Pallmann, P. *et al.* (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1), p. 29. doi: 10.1186/s12916-018-1017-7

Pasanen, T. *et al.* (2019). *What Drives Policy Change in Nepal? A Comparative Analysis*. Overseas Development Institute (ODI), London [online]. Available from: www.econstor.eu/handle/10419/206747 (Accessed 6 October 2021).

Pattyn, V., Molenveld, A. and Befani, B. (2019). Qualitative comparative analysis as an evaluation tool: lessons from an application in development cooperation. *American Journal of Evaluation*, 40(1), pp. 55–74. doi: 10.1177/1098214017710502

Paz-Ybarnegaray, R. and Douthwaite, B. (2017). Outcome evidencing. *American Journal of Evaluation*, 38(2), pp. 275–293. doi: 10.1177/1098214016676573

PMI (2019) *U.S. President's Malaria Initiative FY 2020 Guidance* URL:https://d1u4sg1s9ptc4z.cloudfront.net/uploads/2021/03/pmi-technical-guidance-3.pdf Accessed 2021/10/07

Possebom, V.A. (2017). Free trade zone of Manaus: an impact evaluation using the synthetic control method. *Revista Brasileira de Economia*, 71(2). doi: 10.5935/0034-7140.20170011

Prouty, C., Mohebbi, S. and Zhang, Q. (2018). Socio-technical strategies and behavior change to increase the adoption and sustainability of wastewater resource recovery systems. *Water research*, 137, pp. 107–119. doi: 10.1016/j.watres.2018.03.009

Quast, T. and Gonzalez, F. (2017). Sex work regulation and sexually transmitted infections in Tijuana, Mexico. *Health Economics*, 26(5), pp. 656–670. doi: 10.1002/hec.3339

Ragin, C. (1987). *The Comparative Method: Moving Beyond Qualitative and Quantitative Methods*. University of California, Berkeley, CA.

Ragin, C. (2008). *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. University of Chicago Press.

Reynolds, M. (2014). Equity-focused developmental evaluation using critical systems thinking. *Evaluation*, 20(1), pp. 75–95. doi: 10.1177/1356389013516054

Rihoux, B. and Marx, A. (2013). QCA, 25 Years after 'the Comparative Method'. *Political Research Quarterly*, 66(1), pp. 167–235. doi: 10.1177/1065912912468269

Rogers, P.J. (2008). Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation*, 14(1), pp. 29–48. doi: 10.1177/1356389007084674

Rogers, P.J. (2009). Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation. *Journal of Development Effectiveness*, 1(3), pp. 217–226. doi: 10.1080/19439340903114636

Roig-Tierno, N., Gonzalez-Cruz, T.F. and Llopis-Martinez, J. (2017). An overview of qualitative comparative analysis: a bibliometric analysis. *Journal of Innovation & Knowledge*, 2(1), pp. 15–23. doi: 10.1016/j.jik.2016.12.002

Rothman, K.J. and Greenland, S. (2005). Causation and causal inference in epidemiology. *American journal of public health*, 95 Suppl 1, pp. S144-50. doi: 10.2105/AJPH.2004.059204

Ryckman, T. et al. (2019) 'Impact of Feed the Future initiative on nutrition in children aged less than 5 years in sub-Saharan Africa: difference-in-differences analysis', BMJ (Clinical research ed.), 367, p. l6540. doi: 10.1136/bmj.l6540.

Sarr, B. (2016). Assessing revenue authority Performance in developing countries: a synthetic control approach. *International Journal of Public Administration*, 39(2), pp. 146–156. doi: 10.1080/01900692.2014.1003382

Schneider, C.Q. and Wagemann, C. (2012). *Set-Theoretic Methods for the Social Sciences*, *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge University Press, Cambridge. doi: 10.1017/CBO9781139004244

Schreinemachers, P., Berger, T. and Aune, J.B. (2007). Simulating soil fertility and poverty dynamics in Uganda: a bio-economic multi-agent systems approach. *Ecological Economics*, 64(2), pp. 387–401. doi: 10.1016/j.ecolecon.2007.07.018

Semwanga, A.R., Nakubulwa, S. and Adam, T. (2016). Applying a system dynamics modelling approach to explore policy options for improving neonatal health in Uganda. *Health research policy and systems*, 14(1), p. 35. doi: 10.1186/s12961-016-0101-8

Shiell, A., Hawe, P. and Gold, L. (2008). Complex interventions or complex systems? Implications for health economic evaluation. *BMJ*, 336(7656), pp. 1281–1283. doi: 10.1136/bmj.39569.510521.AD

Sills, E.O. *et al.* (2015). Estimating the impacts of local policy innovation: the synthetic control method applied to tropical deforestation. *PLOS ONE*, 10(7), p. e0132590. doi: 10.1371/journal.pone.0132590

Skivington, K. *et al.* (2021). A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ*, 374, p. n2061. doi: 10.1136/bmj.n2061

Smajgl, A. *et al.* (2015). Assessing the effectiveness of payments for ecosystem services for diversifying rubber in Yunnan, China. *Environmental Modelling & Software*, 69, pp. 187–195. doi: 10.1016/j.envsoft.2015.03.014

Smith, N.R. *et al.* (2018). Agent-Based Models of malaria transmission: a systematic review. *Malaria Journal*, 17(1), p. 299. doi: 10.1186/s12936-018-2442-y

Smith, K., Sridharan, S. and Sattar, S. (2018) The Rockefeller Foundation's Transforming Health Systems Initiative Final Evaluation. The Rockefeller Foundation Monitoring and Evaluation Office. Available at: https://www.rockefellerfoundation.org/wp-content/uploads/Transforming-Health-Systems-Final-Evaluation.pdf.

Springfield Center (2015). *The Operational Guide for the Making Markets Work for the Poor (M4P) Approach* (2nd ed.).

Sterman, J.D. (2002). Systems dynamics modeling: tools for learning in a complex world. *IEEE Engineering Management Review*, 30(1), pp. 42–42. doi: 10.1109/EMR.2002.1022404

Taylor, J.E. *et al.* (2016). Economic impact of refugees. *Proceedings of the National Academy of Sciences*, 113(27), pp. 7449–7453. doi: 10.1073/pnas.1604566113

Taylor, J.E. and Filipski, M.J. (2014). *Beyond Experiments in Development Economics: Local Economy-wide Impact Evaluation*, *OUP Catalogue*. Oxford University Press, Oxford [online]. Available from: https://ideas.repec.org/b/oxp/obooks/9780198707882.html (Accessed 5 October 2021).

Thior, I. *et al.* (2006). Breastfeeding plus Infant Zidovudine Prophylaxis for 6 months vs formula feeding plus Infant Zidovudine for 1 month to reduce mother-to-child HIV transmission in Botswana. *JAMA*, 296(7), p. 794. doi: 10.1001/jama.296.7.794

Thomas, J., O'Mara-Eves, A. and Brunton, G. (2014). Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: a worked example. *Systematic Reviews*, 3(1), p. 67. doi: 10.1186/2046-4053-3-67

Thome, K. *et al.* (2013). Agricultural spillover effects of cash transfers: what does LEWIE have to say? *American Journal of Agricultural Economics*, 95(5), pp. 1338–1344. doi: 10.1093/ajae/aat039

Thorlund, K. *et al.* (2018). Key design considerations for adaptive clinical trials: a primer for clinicians. *BMJ (Clinical research ed.)*, 360, p. k698. doi: 10.1136/bmj.k698

Tougher, S. et al. (2012) 'Effect of the Affordable Medicines Facility - Malaria (AMFm) on the availability, price, and market share of quality-assured artemisinin-based combination therapies in seven countries: A before-and-after analysis of outlet survey data', The Lancet. Elsevier Ltd, 380(9857), pp. 1916–1926. doi: 10.1016/S0140-6736(12)61732-2.

UNFPA-UNICEF (2019) Global Programme to Accelerate Action to End Child Marriage Joint Evaluation. New York: United Nations Children's Fund. Available at: https://www.unfpa.org/sites/default/files/admin-resource/Main_Report_CM_Eval.pdf.

US Department of Health and Human Services, FDA, CDER, C. (2019). *Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry*.

USAID (2011) External Evaluation of the President's Malaria Initiative - Final Report Executive Summary. Washington DC: USAID Global Health Performance Cycle Improvement Project. Available at: https://pdf.usaid.gov/pdf_docs/Pdact332.pdf.

Wagner, A. (1999). Causality in complex systems. *Biology & Philosophy*, 14(1), pp. 83–101. doi: 10.1023/A:1006580900476

Walque, D. de *et al.* (2017). *Looking into the Performance-Based Financing Black Box: Evidence from an Impact Evaluation in the Health Sector in Cameroon*. Policy Research Working Paper. World Bank, Washington DC.

White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation*, 16(2), pp. 153–164. doi: 10.1177/1356389010361562

White, H. (2018). Theory-based systematic reviews. *Journal of Development Effectiveness*, 10(1), pp. 17–38. doi: 10.1080/19439342.2018.1439078

White, H. and Masset, E. (2007). Assessing interventions to improve child nutrition: a theory-based impact evaluation of the Bangladesh Integrated Nutrition Project. *Journal of International Development*, 19(5), pp. 627–652. doi: 10.1002/jid.1344

Widener, M. J. *et al.* (2013). Modeling policy and agricultural decisions in Afghanistan. *GeoJournal*, 78(4), pp. 591–599. doi: 10.1007/s10708-012-9453-y

Windrum, P., Fagiolo, G. and Moneta, A. (2007). *Empirical validation of Agent-Based Models: Alternatives and prospects*. JASSS.

Wossen, T. and Berger, T. (2015). Climate variability, food security and poverty: agent-based assessment of policy options for farm households in Northern Ghana. *Environmental Science & Policy*, 47, pp. 95–107. doi: 10.1016/j.envsci.2014.11.009

Yang, Y. (2019). A narrative review of the use of Agent-Based Modeling in health behavior and behavior intervention. *Translational Behavioral Medicine*, 9(6), pp. 1065–1075. doi: 10.1093/tbm/iby132

Yousafzai, A.K. *et al.* (2015). Parenting skills and emotional availability: an RCT. *Pediatrics*, 135(5), pp. e1247-57. doi: 10.1542/peds.2014-2335

Zazueta, A.E., Le, T.T. and Bahramalian, N. (2021). Development trajectories and complex systems—informed theories of change. *American Journal of Evaluation*, 42(1), pp. 110–129. doi: 10.1177/1098214020947782

Contact us

Centre of Excellence for Development Impact and Learning (CEDIL)
London International Development Centre
36 Gordon Square
WC1H 0PD
www.cedilprogramme.org

Centre of Excellence for Development Impact and Learning

www.cedilprogramme.org

@CEDILProgramme

Centre of Excellence for Development Impact