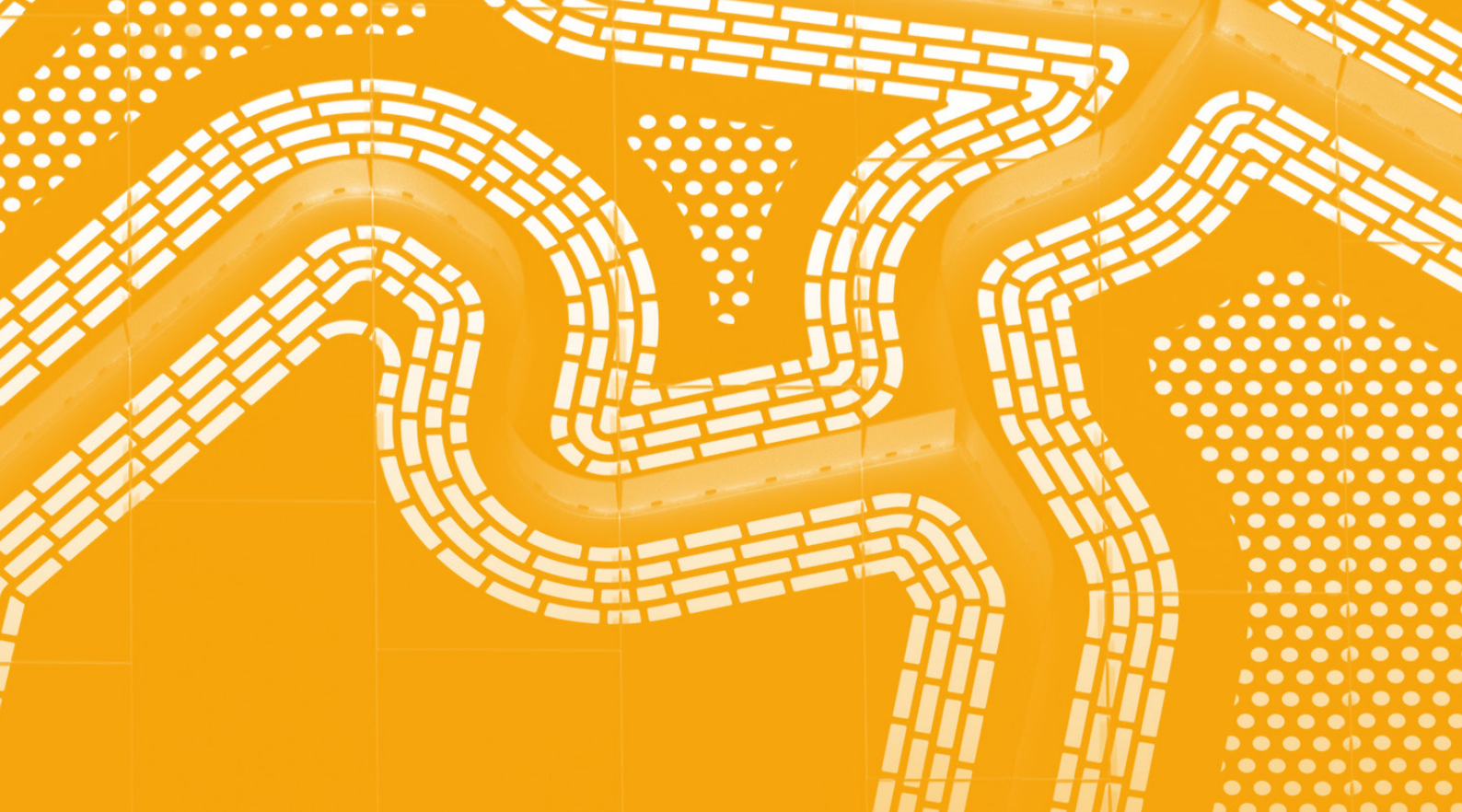


FEBRUARY 2022

CEDIL Methods Brief 7

Evaluating complex interventions: What are appropriate methods?

Centre of Excellence for Development Impact and Learning

Box 1: Highlights

We identify four types of complex development interventions: long causal chain interventions, multicomponent interventions, portfolio interventions, and system-level interventions. These interventions are characterised by multiple activities, multiple outcomes, multiple components, a high level of interconnectedness, and non-linear outcomes.

Complex interventions are difficult to evaluate because effects cannot be unpacked, control groups cannot be built, and outcomes are difficult to observe. Standard impact evaluation methods, such as randomised controlled trials (RCTs) and quasi-experiments, are ill-equipped to evaluate these interventions.

We identify a number of approaches (see the Glossary in Box 2, at the end of this brief) that support the evaluation of different types of complex interventions:

- **Long causal chain interventions** can be evaluated using **pragmatic RCTs** that combine experimental designs with rigorous qualitative process evaluations, or they can be evaluated using **mechanism experiments** that test the impact of specific intervention mechanisms
- **Multicomponent interventions** can be evaluated using experimental approaches when the interventions are piloted: **adaptive trials** allow the selection of the most effective intervention components, while **factorial designs** allow the identification of synergistic effects.
- The impacts over time of large-scale **portfolio interventions** can be evaluated using **synthetic control methods**, while the interactions of these interventions with the context can be understood using **qualitative comparative analysis (QCA)**.
- **System-level interventions** require simulation methods. **Agent-based modelling** and **system dynamics** show the complex impacts of these interventions under different scenarios. However, these methods need to be further refined to allow for causal inference, and need to be validated with real data collected in the field.

Introduction

There is a general perception that the world is becoming increasingly complex and that public policies and interventions are not well-equipped to deal with these increasing levels of complexity. The complexity of social environments makes it difficult to plan interventions, and hard to predict their impacts. This poses a major challenge to evaluators because the evaluator's toolbox is filled with methods that were designed to assess the impacts of relatively simple interventions in stable environments.

Complex interventions are difficult to unpack. They produce multiple outcomes, many of which are measurable only in the long term. The results are difficult to interpret and different stakeholders will value the same results in different ways. Valid control groups for counterfactual analysis are hard to find.

Complex interventions change activities during implementation by adapting to the context, and evaluations of them cannot be set up prospectively.

Despite all these difficulties, we show in this brief that fears that development projects are too complex to be evaluated are somewhat exaggerated. We present a number of evaluation approaches for complex interventions, and use examples to illustrate how they can be applied effectively in evaluations. A few of the methods presented here are not entirely new, but they are totally new to the field of international development. They all need further development and adaptation in order to be effectively employed in the evaluation of complex interventions.

What are complex interventions?

Complex interventions are interventions including multiple activities, operating in different sectors, affecting different population groups, involving different stakeholders, and that are carried out by multiple implementers. They are often implemented at a large scale, they can be modified during implementation, and they may give rise to unexpected impacts and to non-linear outcomes – for example, synergies.

All development interventions are complex to some degree. Rather than classifying interventions as simple or complex, we draw on multidisciplinary literature and on the practice of development interventions to define four typical complex interventions: interventions with long causal chains; multicomponent interventions; portfolio interventions; and system-level interventions. No single real-world intervention will perfectly fit each type, and most interventions will include elements of different types. The typology is useful for evaluators and commissioners to identify the predominant complexity of each intervention and to map out evaluation methods that are appropriate for each intervention type.

Interventions with long causal chains are designed as sequences of causal steps, each of which requires influencing agents to behave in a certain way. Consider, for example, *Educate!* a skills training programme implemented by a non-governmental organisation (NGO) in Kenya, Uganda, and Rwanda, which is being evaluated by a CEDIL-funded team. The programme imparts skills to secondary school students, and seeks to improve their educational attainment, living conditions, and empowerment. The first step in the results chain of the programme involves the NGO persuading a school to partner with them. This leads to training events in schools, which in turn lead to an increase in skills. In the long term, this will increase students' employability and will help them obtain a job. Factors potentially supporting, or derailing, the implementation of the programme can be identified at each step of the chain. For example, students will only gain skills if the programme is implemented properly, the materials are made available, the school prioritises the lessons, and so on. Later steps in the theory of change may be frustrated or enabled by features of the economic and social environment. To sum up, an intervention becomes more complex the longer the causal chain and the larger the number of supporting factors required at each step.

Multicomponent interventions include several activities that are typically expected to produce synergies. They are implemented at the local level, such as in a community or region, and can be carried out within a sector or can cover several sectoral areas, such as education, health, and livelihoods. Project designers include activities the outputs of which interact and mutually reinforce each other to produce an impact that is larger than the sum of the impacts that would be produced by each activity if implemented separately.

Multicomponent anti-poverty programmes are examples of this approach. These programmes offer services in agriculture, education, health, and infrastructure, with the goal of lifting households out of poverty through a 'big push'. An example of this approach is the Ultra-Poor Graduation Initiative. This programme, implemented in several countries by the international NGO BRAC, is a package of cash transfers, training, asset transfers, and financial inclusion interventions that is implemented with the goal of breaking the poverty trap.

Portfolio interventions are projects that are bundled together under a wider umbrella. The projects in the portfolio are implemented across sectors and across different populations; they are implemented at a large scale and over a long period of time, and there is often an expectation that they will positively interact with each other. They have high-level goals which are broadly defined, often in relation to a specific thematic area, such as empowerment, economic development, or climate change. Since they involve many stakeholders and different intervention areas, they tend to be flexible and are adapted to specific contexts. The coordination of the various projects sometimes requires an overarching organisational structure.

Feed the Future is an example of a cross-country portfolio intervention: it is an initiative developed by the US Department of State and coordinated by the United States Agency for International Development's (USAID's) Bureau for Food Security to 'combat global hunger, poverty and malnutrition'. Under this programme, USAID country offices have developed multiyear strategies that include the deployment of a variety of interventions to boost agricultural productivity and improve nutrition, particularly for women and children. 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.

System-level interventions adopt a macro perspective and are not focused on the delivery of specific services or goods to households and individuals but rather set themselves the ambitious goal of changing how the whole system operates, be it the 'education system', the 'health system', or the 'market system'. The change is often non-linear and involves tipping points and far-reaching transitions. These projects target different population groups at different levels in various ways.

They do not have a pre-specified theory of change, as the characteristics of the interventions are developed during project operations, as implementers learn how the intervention works.

The *Making Markets Work for the Poor* (M4P approach) is a system-level intervention that is based on the idea that the poor depend on market systems for their livelihoods, and that improving and changing market systems is key to poverty reduction. In the intervention, product markets are seen as being embedded within a system that includes other markets and supporting factors, such as education, health, water, and sanitation. In this view, all elements of the system are interconnected, and changing just one component is not feasible or even desirable. 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 system is characterised by trade-offs and interventions are implemented to remove those constraints that prevent the efficient operation of the whole system.

What are suitable methods for evaluating complex interventions?

Here we provide some practical pointers about the choice of appropriate methods for evaluating complex interventions, using several examples from the available literature. In doing so, we focus on methods that address causality through counterfactual analysis: either explicitly, by comparing changes in outcomes in a project and in a control group, or hypothetically, by simulating the effects of different interventions.

Interventions with long causal chains can be evaluated using existing methods. Most long causal chain interventions can be unpacked in the form of results chains, and the project effects, if data are available, can be estimated at each step of the chain using experimental or quasi-experimental methods. When the results chains are very complex, researchers can use **pragmatic randomised controlled trials (RCTs).** In a pragmatic RCT, a randomised experiment is conducted alongside a rigorous qualitative process evaluation that analyses the fidelity of the intervention to the project design, and that helps to map out the outcomes of the interventions along the project causal chain. In a pragmatic RCT, the researchers can hypothesise and test the operation of project mechanisms and their interactions with the context.

In some cases, we know from previous studies how the various steps in the causal chain are related but we are uncertain about the mechanism behind one step, or a few steps, in the chain. For example, Ludwig *et al.* (2011) discuss the impact of police patrols against illegal guns in high-crime areas on obesity. The causal chain runs from crime, to psychological stress, and to resulting obesity. Suppose previous work has already demonstrated the impact of police patrols on crime, and also the

impact of crime reduction on stress, but the impact of stress on obesity is unknown. In these cases, a **mechanism experiment** could be used to test the weakest link (in terms of knowledge) in the chain. Researchers could set up an experiment to assess the impact of stress reduction on obesity. A full evaluation of police patrols is not needed, and its impact can be inferred from the impacts that are separately observed at each step of the chain.

Experimental evaluations of multicomponent interventions should be designed in an adaptive way. This is the approach adopted by adaptive trials. In an adaptive trial, the researchers modify the design of the study at various stages of the experiment by successively selecting those intervention components that are more promising. **Adaptive trials** are therefore most useful when interventions can be implemented in many different ways and we do not know which ones are more effective. For example, in an adaptive trial evaluating the promotion of job-seeking among Syrian refugees (Caria *et al*., 2020), the subjects were randomly allocated to four groups: a group receiving an unconditional cash transfer; a group receiving coaching for job interviews; a group receiving a calendar and a job search instruction video; and a control group receiving an information flyer. After six weeks, the subjects were interviewed and were reassigned to the most effective treatment based on employment outcomes up to that point. In this particular case, the three interventions were equally effective in promoting employment, but in an alternative scenario the implementers could have retained the most effective intervention only.

Synergies between project components are best analysed using factorial experiments when piloting interventions. In a **factorial experiment**, many interventions and their interactions are compared to each other. Only a few subjects are assigned to each treatment arm, and the goal of the experiment is not to assess the effectiveness of each intervention but to explore which ones are more promising and deserving of further investigation. For example, a factorial experiment was designed to test the impact of various combinations of parenting interventions in Macedonia, Moldova, and Romania (Lachman *et al*., 2019): a parenting session (short and long); a parenting engagement component (either basic – providing childcare, transportation, and a snack – or *enhanced* – providing food, a raffle prize, cash rewards, and text reminders); and a facilitator supervision component (either *heavy* – five sessions with a trained coach – or *light* – with the coach on demand). Typically, a factorial experiment is able to identify which components are more effective and which ones shows synergistic effects. However, the sample size is usually not sufficiently large to establish the statistical significance of synergistic effects. Hence, in a second phase, the researchers will conduct a randomised evaluation of those interventions and synergies that were found to be promising in the factorial trial.

Synthetic control methods can be used to evaluate portfolio interventions if the goal is simply to assess whether the intervention was effective or not. Synthetic control methods compare the trends of key outcomes in the intervention area to trends observed in similar areas. For example, in a study of the impact of the economic reunification of Germany, the gross domestic product (GDP)

of Germany was compared to a synthetic GDP consisting of a weighted average of the GDPs of Austria, Japan, the Netherlands, Switzerland, and the United States. Synthetic control methods have been used to assess the impact of legislation on tobacco control, health system reforms, changes in taxation, and other wide-ranging policies. Although we were not able to find specific examples of evaluations of portfolio interventions using synthetic controls, it is easy to see how the method could be easily applied to these cases. One limitation in using this approach is that, while country-level data are relatively easy to obtain, long regional time series for the evaluation of specific interventions are more difficult to find.

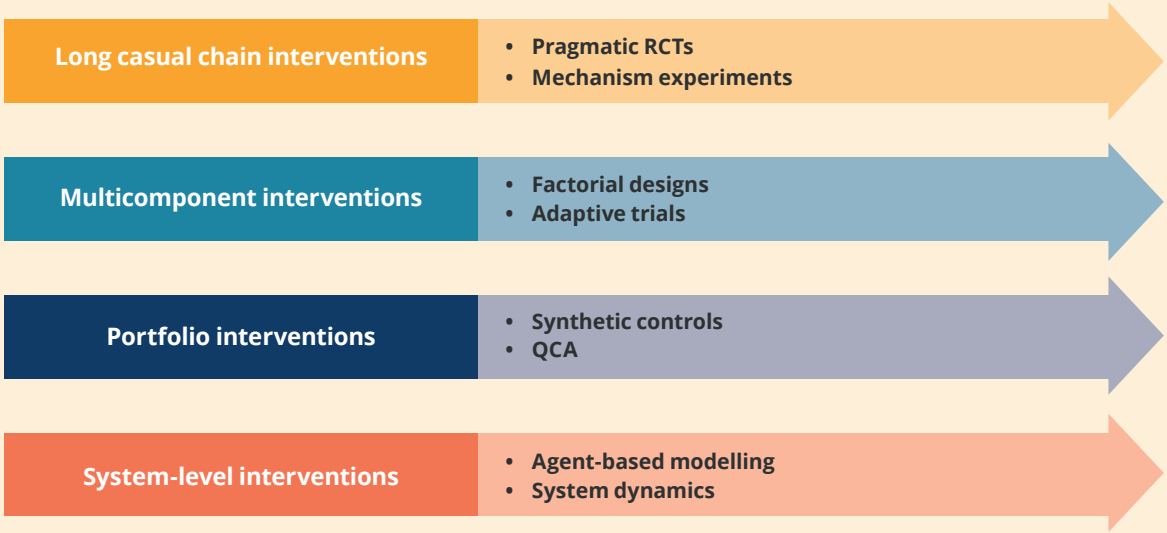
If the goal of the evaluation is to understand what contextual factors are associated with the success of a portfolio of interventions, then qualitative comparative analysis (QCA) is a better approach. QCA is a method of comparative case analysis which identifies configurations of factors that are associated with the success of an intervention. QCA works better where there are few observations, and relies on logical operators rather than statistics. As an example, in an evaluation of irrigation projects in Nepal, a QCA showed that irrigation investments increased agricultural productivity in a sustainable way only when conducted in the presence of farmers' collective action, in the presence of local leadership, and in the absence of external assistance (Lam and Ostrom, 2010). Although we were not able to find examples of applications of this approach to the evaluation of portfolio interventions, it is easy to see how QCA could be applied to the evaluation of these interventions. Researchers have to be aware, however, that the results of QCA are sometimes difficult to interpret, and that they provide weaker evidence of impact in comparison to experimental and quasi-experimental methods.

System-level interventions can only be evaluated using simulation models. System-level interventions are the most difficult to evaluate because evaluations for them cannot be designed prospectively, and because a valid control group cannot be built. Sometimes the success of a system-level intervention crucially depends on the interactions between individuals: for example, in the analysis of the transmission of infectious diseases or of the diffusion of innovations. In these cases, **agent-based modelling** can be used to simulate the interdependencies between people's behaviours. These simulations are

able to predict phenomena that are typical of complex systems, such as tipping points and exponential growth or decay. The simulations are then used to predict the impacts of interventions under different scenarios, thus identifying trade-offs and long-term impacts. For example, in a study of illicit poppy production in Afghanistan (Widener *et al*., 2013) the researchers simulated the influence of insurgents on farmers' decisions, and estimated the impact of two alternative policies to curb poppy production: a subsidy on the production of licit foods, and targeted trafficking blockades at border locations.

Sometimes interventions are implemented in a complex system that is characterised by many interconnected factors and feedback loops. For example, obesity has been described as being affected by hundreds of factors, and of balancing and reinforcing relations between the same factors. In these cases, **system dynamics** approaches can be used. In system dynamics, the researchers build complex diagrams representing the non-linear relationships between factors and simulate the effects of changes occurring in parts of the system. This approach has the advantage of identifying non-linear and unexpected effects in any part of the system. For example, an evaluation of a payment-for-performance scheme in Afghanistan on maternal and child health services (Alonge *et al*., 2017) employed system dynamics after an RCT had failed to identify any impact. The researchers simulated different bonus payments structures against different hypotheses about staff gaming and levels of motivation, and concluded that bonuses were likely to improve service quality when distributed proportionally to health workers' contributions to services.

However, predictions made by simulation models need to be tested and validated with data from the field. System dynamics and agent-based modelling have the advantage of making it possible to assess the impact of hypothetical interventions when a direct observation of the intervention is impossible. However, this comes at the cost of using simplifying assumptions that make their results mostly suggestive or speculative. Both methods can be used to build a better understanding of interventions and of the context in which they operate, but in order to be effectively used in impact evaluations they need further refinement. Simulations need to be integrated with methods of causal inference, either at the design stage, through counterfactual studies that include populations with and without the intervention, or at the analysis stage, through rigorous validation against the data, and sensitivity analysis.



Conclusions

Complex interventions are implemented at a large scale and absorb a large fraction of development aid budgets, and yet they are rarely rigorously evaluated. Evaluations of complex interventions often consist of weakly designed observational studies, desk reviews, and simple qualitative studies. This is partly because these interventions are difficult to evaluate using standard methods of causal inference.

We show that several approaches have been developed in different disciplines that can be adapted to evaluate different types of complex interventions. These approaches can rarely be used in isolation and should be combined in order to answer different types of questions. The majority of these approaches are very rigorous and can be employed at limited cost. They are currently underutilised and it is surprising that they are not more widely used.

Box 2: Glossary of methods for evaluating complex interventions

Pragmatic RCTs are RCTs that are conducted alongside qualitative studies analysing fidelity in the implementation of the intervention. The process evaluation informs the trial by providing testable hypotheses regarding how the intervention operates and under what circumstances. Many RCTs are today conducted alongside qualitative evaluations in order to answer a wider range of questions than simply testing whether the intervention was effective or not effective.

Mechanism experiments are experiments that do not test a policy or evaluate an intervention but that simply test a single causal link of the results chain. They are designed to understand the mechanism operating behind poorly understood causal links. Efficacy trials testing the effectiveness of drugs under controlled circumstances can be seen as an example of mechanism experiments.

Factorial designs were developed by Ronald Fisher and other statisticians alongside randomised experiments. They have found many applications in engineering, particularly in the process of optimising manufactured goods, but have been largely neglected by other disciplines. They are sometimes confused with multi-arm trials, but their goals, characteristics, and interpretation are different. In factorial experiments, the effects of many treatments, and of combinations of treatments, are compared to each other in an exploratory fashion in order to select those treatments and combinations that appear to be more effective.

Adaptive trials were developed in the biostatistics literature of the 1960s and 1970s. Much of this literature has remained entirely theoretical and empirical applications are limited to drug testing by the pharmaceutical industry. Adaptive trials are flexible RCTs in which the researchers can change the design of the study based on the information collected in the early stages of the experiment. In this way, researchers can, for example, drop the experimentation of ineffective treatments, reallocate subjects to the most effective treatments, change the sample size of the study, or even change the research questions.

Synthetic controls are a method developed by Alberto Abadie and co-authors in the 2000s which has found many applications in political science and economics. The method was developed to evaluate the impact of large interventions and policies on single aggregate units, such as countries, states, or regions. The method evaluates the impact of policies by comparing time series of key outcome variables in the treated unit to a 'synthetic' time series that is a weighted average of trends in similar comparison units. Synthetic control methods are applied to small datasets and require a deep knowledge of the cases considered.

QCA is a research method developed by Charles Ragin and co-authors in a series of books in the 1990s. It has been widely used in political science, sociology, and business and management studies. The method is grounded in the notion that effects are not produced by a single cause but by combinations of multiple causes. It uses logic operators of necessity and sufficiency to identify combinations of characteristics of interventions, and of the context, that are associated with successful or unsuccessful outcomes. The method is best applied to small datasets and requires a deep knowledge of the cases considered.

Agent-based modelling has a long history and has been used in various disciplines. It consists of computer simulations of interactions between agents that generate complex patterns. The reproduction of the complex patterns produced by flocking birds is a classic example of how the approach is able to model extremely complex phenomena starting from few and simple behavioural rules. The approach became popular in the social sciences for simulating 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. However, its application to the evaluation of interventions has been very limited.

System dynamics, like agent-based modelling, was designed to understand the operation of complex systems. It was developed in the 1950s through the work of Jay Forrester and was famously applied in the influential *Limits to growth* report commissioned by the Club of Rome in 1972. In system dynamics, all elements of a system are interconnected and the connections are characterised by feedback loops. Concretely, system dynamics research employs two types of tools: causal loop diagrams and feedback models. The approach has not been widely employed in evaluation and it is best considered as a method that invites researchers to adopt a system perspective by delineating the nature of feedbacks loops and by identifying the ramified consequences of different courses of action.

References

Alonge, O., Lin, S., Igusa, T., and Peters, D. H. (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.

Caria, S.,Gordon, G., Kasy, M., Quinn, S., Shami, S., and Teytelboym, A. (2020) 'An adaptive targeted field experiment: Job search assistance for refugees in Jordan', *CESifo Working Paper*.

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

Lachman, J. M., Heinrichs, N., Jansen, E., Brühl, A., Taut, D., Fang, X., Gardner, F. 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.

Ludwig, J., Kling, J. R. and Mullainathan, S. (2011) 'Mechanism experiments and policy evaluations', *Journal of Economic Perspectives* 25(3), pp. 17–38. doi: 10.1257/JEP.25.3.17.

Widener, M. J., Bar-Yam, Y., Gros, A., Metcalf, S. S., and Bar-Yam, Y. (2013) 'Modeling policy and agricultural decisions in Afghanistan', *GeoJournal* 78(4), pp. 591–599. doi: 10.1007/s10708-012-9453-y.

About this brief

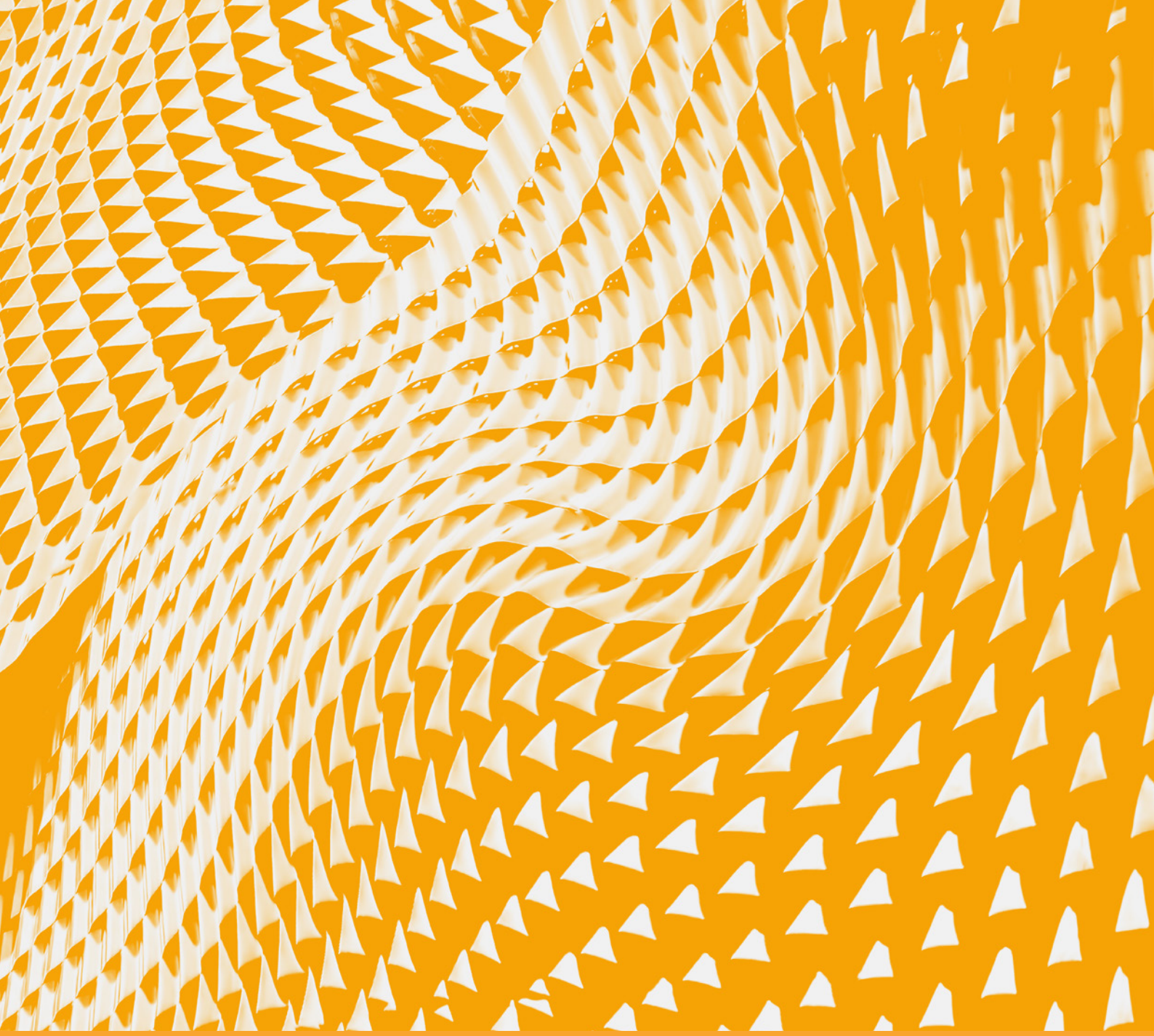
This brief has been prepared by Edoardo Masset. It is primarily based on Masset, E., Shrestha, S. and Juden, M. (2021) 'Evaluating complex interventions in international development', *CEDIL Methods Working Paper 6*. Centre of Excellence for Development Impact and Learning (CEDIL), London and Oxford. Available from: https://doi.org/10.51744/CMWP6

Suggested citation

Masset, E. (2022). 'Evaluating complex interventions: What are appropriate methods?', *CEDIL Methods Brief 7*. Centre of Excellence for Development Impact and Learning (CEDIL), London and Oxford. Available from: https://doi.org/10.51744/CMB7

Photo credits

p.3: Jessica Lea, DFID (now FCDO), P.4: Kyaw Win Khaing, WorldFish, p.5: C.Schubert, p.6: Russell Watkins, DFID (now FCDO), p.7: Nabin Baral, watt solar



@CEDILProgramme

Centre of Excellence for Development Impact

Centre of Excellence for Development Impact and Learning

UKaid from the British people

About CEDIL

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

For more information on CEDIL, contact us at **cedil@opml.co.uk**
or visit our website **www.cedilprogramme.org**