Measuring the Hard-to-Measure in Development: Dimensions, Measurement Challenges, and Responses

Anne L. Buffardi, Tiina Pasanen, and Simon Hearn

1 Introduction

As set out in the 2030 Agenda, sustainable development is universal, integrated, and indivisible, and balances economic, social, and environmental dimensions. Initiatives that aim to improve development attempt to address entrenched issues, where previous efforts have been insufficient or adequate responses are not known. They operate under conditions of uncertainty and complexity. Increasingly they involve multi-component or package interventions, delivered by and relating to diverse sets of stakeholders pursuing different, sometimes competing interests, and working in shifting contexts. Each of these factors carry implications for measurement and pose distinct threats to validity and reliability.

In recent years, there has been a greater acknowledgement of these challenges and the need for robust mixed methods and more flexible, adaptive approaches. At the same time, how best to gather, combine, and interpret multiple sources of evidence covering multi-dimensional aspects of economic and social development, and communicate and use this evidence to guide policy and programming remains a formidable challenge. As applied researchers attempting to gather credible evidence on the effects and trajectories of efforts to improve development, we must continually ask not only "Are we measuring it right?", but also, more fundamentally, "Are we measuring the right thing?"

Based on common challenges that arose through development initiatives in a variety of contexts, this contribution explores four hard-to-measure dimensions of development. In particular, we discuss abstract, multi-dimensional *concepts, processes, and issues*; challenging *settings* where there are unpredictable, sudden, or frequent shifts in the environment; multiple, uncertain *pathways of change*; and multi-layer *implementing structures* such as cross-sector partnerships or regional/

A. L. Buffardi (✉) · T. Pasanen (✉) · S. Hearn (✉)
Overseas Development Institute (ODI), London, UK
e-mail: a.buffardi@odi.org.uk; t.pasanen@odi.org.uk; s.hearn@odi.org.uk

P. Dobrescu (ed.), *Development in Turbulent Times*,
https://doi.org/10.1007/978-3-030-11361-2_3

national/subnational arrangements. These dimensions pose specific difficulties related to *what* is measured and *where*, the assessment of *how and why* changes took place, and *who* or what is the unit of analysis. In addition to these technical, methodological aspects, relational and political factors also have implications for measurement, even if the issue area, setting, pathway of change, and implementing structure are not inherently difficult.

Distinguishing among these four hard-to-measure dimensions is a fundamental first step in minimising potential threats to validity and reliability. To inform this analysis and discussion, we draw on work from policy evaluation and international development, where debates about measurement have featured prominently in recent years.

Scholarship on evidence-informed decision-making has highlighted the importance of the nature of the evidence—its quality, credibility, and relevance—as well as individual, interpersonal, organisational, inter-organisational, and broader contextual factors that can affect the extent to which evidence is used and how (Jones et al. 2012; Bossuyt et al. 2014; Punton 2016). In this contribution, we focus on measurement challenges that affect the credibility of evidence, the sources of information that can help inform development policy and practice. We acknowledge, however, that just as important is the way a particular piece of evidence is integrated with other sources, and how different stakeholders, including decision-makers, are engaged throughout the process.

2 Disconnect Between Research and Reality

There is widespread recognition that development is about multi-dimensional, sustained system-wide changes. The EU, for example, supports programmes on agriculture and rural development, employment and social inclusion, and regional and urban development. The EC Directorate-General for International Cooperation and Development (2017) aims to reduce poverty, ensure sustainable development, and promote democracy, peace, and security. Each of these areas are possible to measure—poverty, inequality, social inclusion, conflict—but represent reversible outcomes that are affected by many factors, which may include but are certainly not limited to a single government ministry or an EU development grant.

The mismatch between expectations and what is plausible to achieve and attribute to a single source can frustrate rather than enhance accountability among those designing, delivering, and intended to benefit from development policies. While these tensions often surface through measurement processes, they reflect more fundamental matters, discussed further in the final section.

The bounded nature of measurement poses additional challenges in trying to assess multi-dimensional, sustained system-wide changes. Like development policies and programmes themselves, measurement is conducted within time and budget constraints. In order to ensure adequate internal validity, account for alternative explanations for change, and strengthen causal claims, research and evaluation must focus on a limited set of questions in order to investigate multiple potential explanatory factors in sufficient depth, comparing instances of presence and absence. Measuring development tends to focus on project-specific changes over relatively short time frames, or very broad aggregates that mask differences across subpopulations and regions. Therefore, the broad, interconnected, and long-term nature of social and economic change processes, and the necessarily bounded nature of development initiatives and their measurement, can be difficult to reconcile.

2.1 Complicated, Complex, and Hard to Measure

Indeed, these challenges have re-energised the discourse on complexity, which has been prominent in the field of international development and in cross-government initiatives in some EU countries, and is widely relevant across a range of contexts (Eoyang and Berkas 1998; Glouberman and Zimmerman 2002; Kurz and Snowden 2003; Ramalingam et al. 2008; Mowles et al. 2008; Hall and Clark 2010; Rogers 2011; Hummelbrunner and Jones 2013a, b; Mowles 2014; Copestake 2014; Matthews 2016; Root et al. 2015; Buffardi 2016). Scholars have drawn insights from complexity theory, characterising key elements in different ways (Fig. 1). They note the importance of clarifying which *aspects* of a development initiative have these features, rather than characterising the initiative in its entirety as complex (Rogers 2011; Woolcock 2013; Yin 2013).

As attention to complexity has heightened, there is a risk that the term is misused and overapplied to situations or aspects of a programme that do not fit these characteristics but that may nevertheless be difficult or challenging. As Peersman et al. (2016) note, what is complex is not just very, very complicated; rather, it is characterised by its dynamic and emergent nature, which requires ongoing knowledge generation to gauge what is working given current conditions and what is the best way forward. These questions imply a different focus than with complicated (rather than complex) aspects of interventions, which ask what works more and less well for whom and under what conditions.

In working with development initiatives—most of which have elements that are fairly straightforward, complicated, and uncertain—we have found that measurement challenges are just as often complicated as they are complex; and that these terms are applied inconsistently, sometimes causing confusion for decision-makers and programme staff. Therefore, we use the term “hard-to-measure” to avoid this confusion and focus on the common challenges that emerged in working with large development initiatives. These challenges fall along four dimensions: abstract, multi-dimensional concepts, processes, and issues; challenging settings; multiple, uncertain pathways of change; and multi-layer implementing structures. We explore each in turn in the next section.

For some areas with well-established assessment indicators, measurement can be problematic because of the mismatch between expectations and what is plausible and feasible to observe over what periods of time, rather than because of technical, Core features of complex programmes:

1. *Distributed capacities*, when skills, resources and actions are dispersed across many actors, institutions and geographies, whose joint interaction is required to address a problem
2. *Goal divergence,* when these actors have different perspectives about the problem and/or how to best address it
3. *Uncertainty*, incomplete knowledge or understanding about how to achieve desired outcomes in a particular context

Ramalingam et al. 2008, Jones 2011

Causal density, based on:

- *Intensity of transactions* among individuals
- *Level of discretion* of implementation staff
- *Contentiousness*, pressure to do something other than implement a solution
- Existence of *known solutions*

Woolcock 2013

- *Confidence of causality:* the extent to which the problem and the effectiveness of the response is understood
- *Confidence in context*: the extent to which the political and operational contexts are understood and can be influenced

IDS *&* USAID 2015, Valters et al. 2016

- *Problem-related complexity,* one with multiple elements, variability in the environmental and socio-demographic characteristics of the area, spread and scale of the problem, level of unpredictability
- *Policy/response related complexity*, multiple components, multiple stakeholders, degree of flexibility or tailoring during implementation, spread and scale of the response, competing or interacting policies
- *Impact-related complexity*, multiple expected outcomes, unexpected impacts, interactions between policy components, difficulty in causal attribution, time scales over which impacts might occur, availability of information relating to impacts

Centre for the Evaluation of Complexity Across the Nexus (Sheate et al. 2016)

Fig. 1 Different characterisations of complexity, as applied to development

methodological challenges themselves. We also acknowledge and briefly discuss these situations in the final section.

3 Distinguishing the Hard-to-Measure Aspects of Development: Four Dimensions

Each hard-to-measure dimension corresponds to the what, where, how, why, and who of implementation and measurement. Each dimension faces distinct threats to reliability and internal validity, the foundational elements of any measurement exercise: whether repeated investigation would yield the same findings, and the credibility of the results, whether the investigation measures what it intends to measure.

Internal validity can be compromised by a number of factors, including *history effects*, events that occur in the external environment between measurement time points; *maturation*, processes within the individual operating as a function of the passage of time, not related to particular events (i.e. ageing); *testing*, effects of the measurement exercise itself; *instrumentation*, when measurement tools or processes vary; *selection*, when individuals choose to take part in an intervention or are identified in non-systematic ways; and *attrition*, when people leave the programme before it is completed (Campbell and Stanley 1963). More robust study and instrument designs can minimise these threats, and so may be particularly warranted for specific hard-to-measure dimensions.

Table 1 presents each dimension with an example and the primary measurement challenges and risk associated with it. After examining each in turn, we discuss interactions and commonalities among the four dimensions, including additional biases which affect measurement and implications for external validity.

Table 1 Hard-to-measure dimensions

Dimension	Explanation and example	Measurement challenge and risk
Abstract, multi-dimensional *concepts, processes, and issues* (what)	Intangible, often unobservable concepts for which proxy indicators must be used: development, social inclusion, empowerment, strengthened partnerships, and institutions	• Construct validity—does the indicator represent the concept it's trying to measure? Risk of measuring the wrong thing
Challenging *settings* (where)	Unpredictable, sudden, and/or frequent shifts in the implementing environment	• Lack of data • Reliability • Instrumentation threat to validity • Higher attrition, changes in the profile of participants/ sample • Higher likelihood of history effects Risk of measuring different things and/or in different ways
Multiple, uncertain *pathways of change* (how and why)	Multiple, interacting, and/or unforeseen trajectories of change leading to multiple, unforeseen outcomes	• Causal attribution, given equifinality and multifinality • Omitted variable bias Risk of misattribution, neglecting key independent or dependent variables
Multi-layer *implementing structures* (who)	Multi-organisation, multi-site, multi-layered structures: cross-sector, cross-department, regional/ national/subnational initiatives	• Unit of analysis Risk of over-aggregation, conflating delivery mechanism with the intervention

3.1 Abstract, Multi-Dimensional Concepts, Processes, and Issues

The first hard-to-measure dimension relates to abstract, multi-dimensional concepts, processes, and issues—*what* precisely is being measured. This relates to Jones' complexity feature of goal divergence, when actors have different perceptions of the problem. The very concept of development exemplifies this dimension. What exactly does sustainable development entail? What types of inequality are most problematic? Concepts like social inclusion, empowerment, accountability, and strengthened partnerships or institutions reflect intangible concepts that may be conceived of and interpreted differently. Therefore, unlike a child's weight, household assets, or agricultural yields, they require the use of proxy indicators to assess their presence, absence, or strength. These proxies themselves may be difficult to observe and so may be more heavily reliant on perceptions rather than measured directly.

For example, Sustainable Development Goal 5, which aims to achieve gender equality and empower all women and girls, uses 14 indicators to measure these concepts, including legal frameworks to enforce non-discrimination, prevalence of intimate partner violence, early marriage, female genital mutilation, unpaid domestic and care work, seats held in national and local government, women in managerial positions, access to information about and decisions regarding sexual and reproductive health and rights, agricultural land ownership, mobile phone ownership, and presence of systems to track and make public allocations for gender equality and women's empowerment (UN-DESA n.d.). The OECD operationalises "better life" as housing, income, jobs, community, education, environment, civic engagement, health, life satisfaction, safety, and work-life balance.

Development programmes often aim to improve a mix of issues that are straightforward and more difficult to measure: increased household income and forest coverage as well as more participatory and inclusive decision-making in natural resource management committees and stronger cross-sector partnerships. The key measurement challenge with abstract, multi-dimensional concepts and issues is construct validity: to what extent does the indicator(s) accurately reflect the underlying theoretical construct or concept? Simply stated, poor construct validity risks measuring the wrong thing.

This measurement challenge can be mitigated by providing clear operational definitions; for example, explicitly stating "by gender equality, this programme is referring to changes in the proportion of women elected to and serving on district councils". For programmes involving many stakeholders, it may require bringing these groups together (i.e. researchers across disciplines, local government officials, programme staff, community members) to discuss their interpretation of the concept. As illustrated above, multi-dimensional concepts are also addressed by using multiple indicators or composite indices to capture different elements of a broader concept. The use of multiple indicators tailored to a particular context may increase construct validity in that study. At the same time, the use of different indicators across programmes or locations limits comparisons and precludes quantitative aggregation like meta-analyses so the broader field may have a less-consolidated evidence base.

3.2 Challenging Settings

The second dimension refers to *where* development policies, programmes, and their measurement are taking place. The economic, political, and/or physical environment may be unpredictable and highly unstable. In some settings, instability may be more predictable in the sense that these shocks may be anticipated but they are frequent and interrupt both programming and measurement. Or, destabilising events may be infrequent and predicted but with rapid onset, such as natural disasters, which leave little time to change activities if contingency plans have not already been established. These settings reflect volatility, the likelihood that important contextual or causal conditions may change quickly (Booth et al. 2016) and are one example of context complexity. This may include shocks as a result of an economic crisis, the intensity and manifestation of which may vary by sector and locale. It could be geographical pockets where migration flows significantly affect the local context, or agricultural areas that are more prone to natural disasters.

Unpredictable, frequent, and/or rapid shifts in the environment affect measurement reliability, whether repeated assessment would find the same results. Rather than measuring the wrong thing, as is the risk with multi-dimensional concepts, challenging settings risk measuring different things and/or in different ways.

If environmental or security concerns prevent access to certain areas or groups, the timing of data collection may be delayed. In some cases, it may not be possible to take multiple measures over time. If these shocks prompt population movement, attrition may be higher than average. The profile of participants or the population sample from which the measures are taken may differ over time. The use of alternative approaches, like remote monitoring techniques, may increase threats to internal validity as a result of instrumentation; that is, observed changes may be a result of the different way data was gathered at different points in time: direct measurement and then remote monitoring. Attempts to minimise these measurement challenges may involve taking more frequent measurements, increasing initial sample sizes, investing more in the follow-up to find displaced participants, and using multiple instruments or assessment measures; for example, using both remote and direct monitoring when access is not limited in the event that it is later restricted.

Measurement within shifting contexts is more likely to suffer from history effects, changes over time in the external context that contribute to or explain the observed outcomes rather than the intervention itself. And, the unique nature of the context may make it difficult to identify suitable comparison groups, which could help to account for history effects.

Other challenges may not be able to be addressed and therefore must be taken into account in the analysis and interpretation. Information may have been destroyed, or access to documents and/or people may be tightly controlled. Even if it is possible to access specific population groups, refugees, for example, they may be reluctant to disclose information (Jones and Pellini 2009; Cramer and Goodhand 2011).

There have been extensive discussions of some of these challenges in the humanitarian sector, which offer important lessons for other contexts facing environmental, political, or economic instability (Waldman 2014; Bush and Duggan 2015; ALNAP 2016).

3.3 Multiple Uncertain Pathways of Change

How and why changes take place—particularly when multiple, interacting, and/or unknown trajectories of change may lead to multiple and unforeseen outcomes—is the dimension discussed most frequently in the literature. Many of the categorisations of complexity relate to this dimension: Woolcock's four aspects of causal density, IDS and USAID causal complexity, and two of Jones' three features of complex programmes: uncertainty about how to achieve desired outcomes and goal divergence regarding how best to address them.

Where a variable is situated along the pathway of change may be uncertain or contested. For example, strengthened partnerships across actors in different sectors may be an intended outcome of a programme and/or it may be seen instrumentally as a way to enable economic and social outcomes for rural farmers.

The key measurement challenge here is causal attribution, given equifinality (multiple pathways) and multifinality (multiple outcomes). With a large number of potential variables and configurations, omitted variable bias may be a concern—the possibility that changes are influenced by factors that are not being measured. Thus, misattribution and oversight of key independent and/or dependent variables are the primary risks.

Different methods attempt to account for multiple independent variables and the interactions between them and improve causal inference. Multivariate regression can include interaction effects. Process tracing systematically investigates alternative explanations for change (Collier 2011). Qualitative comparative analysis tries to identify necessary and sufficient conditions associated with a particular outcome (Befani 2016). Realist evaluation examines context-mechanism-outcome configurations (Westhorp 2014). The relative merits of different methods have been discussed extensively elsewhere (Stern et al. 2012; Stame 2010; White 2010; Chambers et al. 2009); however, there are still relatively few examples of these latter approaches being applied in development.

3.4 Multi-Layer Implementing Structures

The final hard-to-measure dimension is more complicated than complex: multi-organisational, multi-site, multi-layer, often multi-sector implementing structures.

It refers to *who* or what unit(s) are being measured.[1] In this case, it reflects who is delivering the policy or programme and, correspondingly, the unit(s) of data collection, analysis, and reporting. Multi-layer implementing structures reflect Jones' category of distributed capacities, when skills, resources, and actions are dispersed across many actors, institutions, and geographies, whose combined efforts are (thought to be) required to address a problem.

Multi-unit implementing structures appear to be becoming more common. These include cross-sector initiatives and those operating at regional, national, and subnational levels. They could include public-private partnerships or consortium arrangements where multiple organisations work together to deliver a joint EU-funded development project, and multi-project programmes where multiple consortia or projects are grouped together under a wider programme umbrella around a common theme and funding source (Buffardi and Hearn 2016). Numerous programmes may then be nested within broader ministry portfolios or workplans. Other more complicated implementing structures, relative to a single organization or department delivering a single site project, also include networks and coalitions (Hearn and Mendizabal 2011) and regional NGO associations (Davies 2016).

In these types of initiatives, the measurement challenge relates directly to the structure: determining the appropriate unit of collection and analysis. The risk with multi-layer initiatives is over-aggregation, conflating dissimilar units and presenting findings together rather than according to more cohesive and logically or operationally bounded subgroups (Bowman et al. 2013). A large diverse agriculture portfolio with scores or sometimes even hundreds of projects can look broadly at spend rates, the reach and profile of individuals with whom the projects have interacted, and identify particular cases to illustrate different elements of the portfolio and the extent to which those projects contribute to specific outcomes. However, assessing the entire portfolio or a large multi-project programme in aggregate may not be appropriate.

The risk of over-aggregation can be addressed by gathering, analysing, and presenting data according to its multiple nested layers. Clarifying common elements or the purpose of more aggregate structures can help determine the extent to which it is appropriate to use standardised approaches and combine data. Although there has been increasing recognition of the importance of disaggregating data, relative to the other three dimensions, the implications of multi-layer implementing structures are discussed much less often in the development and evaluation literatures.

In our recent experience, challenges related to this dimension have been more prominent than the other three in practice, in part perhaps because measurement challenges are confronted at the outset of implementation when programmes attempt to design a common measurement framework rather than at later phases of the programme when causal pathways are being tested. When development policies and programmes are led by a single government ministry, this dimension is less of an

[1] It is also fundamental for accurate measurement to determine the correct unit of analysis at a project level—whether to gather data on individuals or households, for example.

issue. However, it becomes more salient as more stakeholder groups become involved: multiple departments and agencies at regional, national, and subnational levels, private sector, nongovernmental organisations, farmer associations, and community groups.

3.5 Commonalities and Intersections Among Hard-to-Measure Dimensions

The particular threats to internal validity and reliability highlighted above represent those that are most problematic for each dimension. Measurement efforts associated with all four dimensions may also be affected by other factors, including recall, interviewer, testing, social desirability, and confirmation biases. Incomplete and inconsistent data collection, which is not uncommon with routine monitoring data from development programmes, may produce biased, non-representative findings.

In addition to the threats to internal validity, these four dimensions all have implications for external validity, limiting the extent to which findings can be generalised to other populations and settings. For instance, how abstract, multi-dimensional concepts are interpreted is likely to vary across districts, countries, and population groups. Challenging settings are more likely to affect the delivery of development programmes and directly influence outcomes so findings may be generally relevant to settings that share similar characteristics (i.e. frequent flooding, a weak decentralised government) but it is unlikely that they will be directly applicable. Similarly, the interactions of specific actors and configurations of them may influence delivery and outcomes so findings may differ when programmes are led by a single department or organisation or by different implementing structures and different sets of actors within them. When multiple interacting pathways of change and many outcomes are possible, some pathways and outcomes may be particularly relevant for certain populations and contexts more than others. By their very nature, elements of development programmes that are truly complex (rather than very complicated) are unknowable (Peersman et al. 2016) so specific findings cannot be directly transferred to other situations. However, there may be lessons about processes and adaptation that are more broadly relevant.

Across the four dimensions there may also be areas of overlap and interdependency. Uncertain pathways of change, in particular, may be affected by the other three hard-to-measure areas. Concepts and issues that are less well defined may have more potential trajectories of change, which may also be poorly specified. The roles of individuals, subgroups, and larger umbrella group structures may complicate causal attribution. In addition, history effects that are more likely in challenging settings raise alternative explanations for change that must be taken into account when attempting to assess causality.

4 Discussion and Conclusions

In the last two decades, attention to the measurement and results of government programmes and development initiatives has heightened significantly. There have been lively debates in the field about the relative merits of different methodological approaches and extensive discussions about complexity. At the beginning of the Sustainable Development Goal (SDG) era, we know considerably more about development than we did when the first UN Human Development Report was launched in 1990.

At the same time, formidable challenges remain—both in terms of measurement and, more fundamentally, how to advance the multi-dimensional, sustainable, system-wide changes that development aims to achieve. Identifying what is hard to measure can help specify threats to internal and external validity and reliability so they can be minimised and accounted for in study design, analysis, and interpretation. There is cause for both optimism and vigilance—measurement can be improved to enhance understanding of hard-to-measure areas but researchers must continually question and advance their approaches in order to keep pace with the world we are investigating.

4.1 The Feasibility of More?

We acknowledge that attempts to address the measurement challenges in each hard-to-measure dimension all involve doing *more* of something: using multiple measures to capture a broad concept, gathering data more frequently, with multiple instruments and on larger samples to anticipate shifts in the context, restricted access and higher attrition, gathering and analysing information at each of the multiple layers of complicated implementing structures, and testing many potential pathways of change. The extent to which more monitoring and more comprehensive evaluation is feasible in government and development programmes whose primary aim is implementation may limit the degree to which these challenges can be addressed. For selected conceptual areas, settings, pathways of change, and implementing structures, approaching these challenges through more in-depth substantive research may be more viable than through a relatively small monitoring and evaluation component of a large development initiative.

4.2 Relational and Political Factors Affecting Measurement

Furthermore, although this chapter has focused on measurement challenges, it is just as important to recognise relational and political factors that influence the design, implementation, and evaluation of development programmes. As noted in the introductory chapter, development is far from being solely a "technical" problem. One drawback of the heightened attention to measurement issues has been an overemphasis on technical elements and relative lack of attention to these relational factors. These difficulties often arise as a result of a mismatch between expectations and what is plausible and feasible to observe in a particular time frame—essentially, trying to measure and attribute too much too soon. Pressures to demonstrate "success" and value for money can exacerbate these unrealistic expectations.

Some pathways and outcomes may not necessarily be complicated or complex but may take time before they can be observed. Assessing sustainability, for example, asks to what extent programmes and benefits continued after the programme ended. This question is not inherently difficult but by definition it cannot be assessed until after the programme period.

In addition to pressures to measure sustainability or impacts before they can be observed are situations where programmes attempt to measure too much. Laundry lists of questions and indicators often reflect efforts to be inclusive and covering multiple perspectives and dimensions. However, attempts to answer too many questions result in none being measured in a credible, robust way. Issues of attribution present another challenge—claiming rather than attempting to measure too much. It is rare that a single programme or actor is the only one trying to address a particular issue or working with a group of people.

Relational and political factors are present even in small, relatively straightforward projects but may be intensified in programmes with hard-to-measure elements and manifested in different ways across the four dimensions. By definition, multi-layer implementing structures involve many diverse actors, with different levels of authority and power, who may have different priorities. They may see one another as competitors rather than collaborators and so may be reluctant to share information. Prioritising some questions over others, or using one operational definition of an abstract, multi-dimensional concept rather than another, is an inherently political process. In effect, it validates or elevates certain conceptions and minimises or ignores alternative interpretations. Resolving contested understandings of women's empowerment requires transparent processes of deliberation, not necessarily more or a different type of information.

Similarly, choosing which indicators are included and which are excluded in a national development index involves trade-offs. This is also the case for multiple, uncertain pathways when a long list of potential mediating and outcome variables will need to be bounded. The need to make quick decisions in response to a rapidly shifting environment precludes extensive, inclusive, deliberative processes.

Guidance and evaluation tools exist to help structure processes of stakeholder discussion and prioritisation processes (Peersman et al. 2015; Hearn and Buffardi 2016). However, deciding how development programmes should be judged and resources allocated are potentially contentious processes and there is not a technical fix that can be applied. At the heart of measuring and claiming too much too soon are questions about accountability. Who is responsible to whom, for what and when?

Distinguishing relational political challenges from technical methodological ones and clarifying among specific hard-to-measure dimensions are important because they require different responses to address the underlying problem. They pose distinct threats to measurement validity, reliability, feasibility, and use of evidence that need to be accounted for and minimised to the extent possible. Credible evidence is a foundational component of understanding and improving development—and is indeed a necessary but insufficient condition for evidence-informed decision-making.

References

ALNAP. (2016). *Evaluation of humanitarian action guide: Active Learning Network for Accountability and Performance (ALNAP) in humanitarian action*. London: ALNAP/ODI.

Befani, B. (2016). *Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA)*. Stockholm: EBA.

Booth, D., Harris, D., & Wild, L. (2016). *From political economy analysis to doing development differently: A learning experience*. London: ODI.

Bossuyt, J., Shaxson, L., & Datta, A. (2014). *Assessing the uptake of strategic evaluations in EU development cooperation*. http://ec.europa.eu/europeaid/how/evaluation/evaluation_reports/documents/2014/1331_uptake_strategic_evaluation_annexes_en.pdf

Bowman, K., Buffardi, A. L., & Freeman, S. (2013). *Who interacts with whom, about what and when? Evaluation and learning in multi-issue global networks.* Paper presented at the American Evaluation Association, Washington, DC.

Buffardi, A. L. (2016). *When theory meets reality: Assumptions, feasibility and implications of a complexity-informed approach* (A Methods Lab publication). London: Overseas Development Institute.

Buffardi, A. L., & Hearn, S. (2016). *Multi-project programmes: Functions, forms and implications for evaluation and learning. A Methods Lab publication*. London: ODI.

Bush, K., & Duggan, C. (2015). *Evaluation in the extreme: Research, impact and politics in violently divided societies*. New Delhi: SAGE.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Chambers, R., Karlan, D., Ravallion, M., & Rogers, P. (2009). *Designing impact evaluations: Different perspectives. International Initiative for Impact Evaluation Working Paper 4*. New Delhi: 3ie.

Collier, P. (2011). Understanding process tracing. *Political Sciences and Politics, 44*(4), 823–830.

Copestake, J. (2014). Credible impact evaluation in complex contexts: Confirmatory and exploratory approaches. *Evaluation, 20*(4), 412–427.

Cramer, C., & Goodhand, J. (2011). Hard science or waffly crap? Evidence-based policy versus policy-based evidence in the field of violent conflict. In K. Bayliss, B. Fine, & E. Van Waeyenberge (Eds.), *The political economy of development: The World Bank, neoliberalism and development research*. London: Pluto Press.

Davies, R. (2016). *Evaluating the impact of flexible development interventions using a "loose" theory of change. A Methods Lab publication*. London: ODI.

Eoyang, G. H., & Berkas, T. (1998). *Evaluation in a complex adaptive system: Managing complexity in organizations*. Westport, CT: Quorum Books.

European Commission Directorate-General for International Cooperation and Development (DG DEVCO). (2017). *Mission statement*. Retrieved September 4, 2017, from https://ec.europa.eu/europeaid/mission-statement_en

Glouberman, S., & Zimmerman, B. (2002). *Complicated and complex systems: What would successful reform of Medicare look like?* Toronto: Commission on the Future of Health Care in Canada.

Hall, A., & Clark, N. (2010). What do complex adaptive systems look like and what are the implications for innovation policy? *Journal of International Development, 22*(3), 308–324.

Hearn, S., & Buffardi, A. L. (2016). *What is impact? A Methods Lab publication*. London: ODI.

Hearn, S., & Mendizabal, E. (2011). *Not everything that connects is a network*. London: ODI.

Hummelbrunner, R., & Jones, H. (2013a). *A guide for planning and strategy development in the face of complexity. ODI Background Note*. London: ODI.

Hummelbrunner, R., & Jones, H. (2013b). *A guide to managing in the face of complexity. ODI Working Paper*. London: ODI.

Institute of Development Studies (IDS) and USAID Learning Lab. (2015). *Learning to adapt: Exploring knowledge, information and data for adaptive programmes and policies. Workshop Report*. Brighton: IDS.

Jones, H. (2011). *Taking responsibility for complexity*. London: ODI.

Jones, N., & Pellini, A. (2009). *Evidence-informed policy in post-conflict contexts: Nepal, Peru and Serbia*. London: ODI.

Jones, H., Jones, N. A., Shaxson, L., & Walker, D. (2012). *Knowledge, policy and power in international development: A practical guide*. Bristol: Policy Press.

Kurz, C., & Snowden, D. (2003). The new dynamics of strategy: Sensemaking in a complex world. *IBM Systems Journal, 42*(3), 462–483.

Matthews, M. (2016). How better methods for coping with uncertainty and ambiguity can strengthen government – Civil society collaboration. In G. Carey, K. Landvogt, & J. Barraket (Eds.), *Designing and implementing public policy: Cross-sectoral debates*. London: Routledge.

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

Mowles, C., Stacey, R., & Griffin, D. (2008). What contribution can insights from the complexity sciences make to the thinking and practice of development management? *Journal of International Development, 20*(6), 804–820.

Peersman, G., Guijt, I., & Pasanen, T. (2015). *Evaluability assessment for impact evaluation: Guidance, checklists and decision support. A Methods Lab publication*. London: ODI.

Peersman, G., Rogers, P., Guijt, I., Hearn, S., Pasanen, T., & Buffardi, A. L. (2016). *When and how to develop an impact-oriented monitoring and evaluation system. A Methods Lab publication*. London: ODI.

Punton, M. (2016). *Building Capacity for the Uptake of Research Evidence (BCURE) literature review: How can capacity development promote evidence-informed policymaking?* Brighton: ITAD.

Ramalingam, B., Jones, H., Reba, T., & Young, J. (2008). *Exploring the science of complexity: Ideas and implications for development and humanitarian efforts. ODI Working Paper 285*. London: ODI.

Rogers, P. J. (2011). Implications of complicated and complex characteristics for key tasks in evaluation. In K. Forss, M. Marra, & R. Schwartz (Eds.), *Evaluating the complex: Attribution, contribution, and beyond* (Comparative Policy Evaluation series) (Vol. 18, pp. 33–52). New Brunswick, NJ: Transaction.

Root, H., Jones, H., & Wild, L. (2015). *Managing complexity and uncertainty in development policy and practice*. London: ODI.

Sheate, W.R., Twigger-Ross, C., Papadopoulou, L., Sadauskis, R., White, O., Orr, P., Phillips, P., & Eales, R. (2016) *Learning lessons for evaluating complexity at the nexus: A meta-evaluation of CEP projects*. Final Report to CECAN. https://doi.org/10.13140/RG.2.2.21468.18565.

Stame, N. (2010). What doesn't work? Three failures, many answers. *Evaluation, 16*(4), 371–387.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations. Report of a study commissioned by the Department for International Development. DFID Working Paper 38.* (Vol. DFID), London.

UN-DESA and Sustainable Development Goal 5. Retrieved from http://www.un.org/sustainabledevelopment/gender-equality/

Valters, C., Cummings, C., & Nixon, H. (2016). *Putting learning at the centre: Adaptive development programming in practice*. London: ODI.

Waldman, T. (2014). The use of statebuilding research in fragile contexts: Evidence from British policymaking in Afghanistan, Nepal and Sierra Leone. *Journal of Intervention and Statebuilding, 8*, 149–172.

Westhorp, G. (2014). *Realist impact evaluation: An introduction. A Methods Lab publication*. London: ODI.

White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation, 16*(2), 153–164.

Woolcock, M. (2013). Using case studies to explore the external validity of "complex" development interventions. *Evaluation, 19*(3), 229–248.

Yin, R. K. (2013). Validity and generalization in future case study evaluations. *Evaluation, 19*(3), 321–332.