3. Dealing with Complexity in an Increasingly Interconnected World

MICHAEL WOOLCOCK
World Bank and Harvard University[162]

In this second decade of the twenty-first century, there are ever-rising expectations and demands on the international development community. We have dramatically “raised the bar’” ourselves with the passage of the Sustainable Development Goals (SDGs), which commit 193 governments around the world, rich and poor alike, to achieving, by 2030, outcomes ranging from eliminating poverty and hunger to providing quality health care, justice and education for all (at all ages!). The SDGs are not merely “upgrades” from the eight Millennium Development Goals that preceded them but qualitative shifts in scale, scope and complexity; for some they may be “noble”, “inclusive” and commendably “ambitious” but from a public administration and political theory perspective there is a reasonable concern that they establish expectations in certain key domains (more on this below) that the prevailing implementation capabilities of most non-Organisation for Economic Co-operation and Development (OECD) governments demonstrably—at least at current levels—surely cannot realize (Andrews et al 2017). For present purposes, moreover, they place enormous pressure on corresponding evaluation teams, who in due course will be called upon to assess whether indeed the policies and programmes of all 193 governments have yielded outcomes that are “on track” to meeting the 232 “indicators” by which success on the SDGs will be determined.

Beyond the community of development professionals, the world itself is generating demands—whether through stronger citizen “voice” demanding improvements in the quality of service delivery (e.g., in the Middle East; see Brixi et al 2015) or domestic political events whose effects radiate regionally, even globally (e.g., refugee crises, migration flows, trade disputes, civil wars)—that stretch the implementation capability of even the most solvent and experienced public sectors, let alone those whose budgets are threadbare, whose legitimacy is perhaps questionable and who have little collective experience at managing large-scale, deeply complex policy challenges. Thus, on both the “supply” and “demand” sides, governments everywhere face rising implementation challenges borne of interlocking events and expanding expectations, but an abiding concern that their delivery systems for managing

[162] The views expressed in this speech (and accompanying summary text) are those of the author alone, and should not be attributed to the World Bank, its executive directors or the countries they represent.

them—and the corresponding evaluation tools needed to monitor and assess their effectiveness—may not be up to the task, in so doing risking becoming themselves part of the problem (rather than part of the solution). What to do?

Faced with such challenges, one instinctive response has been to fortify the empirical foundation on which development decisions are made. In effect, the claim is that by providing skeptical or risk-averse policymakers with "rigorous evidence" that certain development interventions do in fact "work", the burden will be lighter upon those tasked with responding to today's global challenges. In the face of deep uncertainty, it can be correspondingly reassuring when bona fide development "experts" provide what seems to be compelling evidence regarding the efficacy of certain "tools" and "best practices". While more and better evidence is always a good thing and recognizing the importance of helping decision makers think systematically about their policy options, the very definition of complex development challenges is that neither the core underlying problem nor the appropriate solution is clear, at least ex ante. A hammer is great if my problem is a nail, but mostly useless if it turns out that what I actually need is a screwdriver. In complex circumstances, therefore, we don't need experts selling us hammers; we need partners who can help us nominate and prioritize our problems, shaping them into manageable sizes so that plausible next steps can be discerned. To respond to these problems, in all their almost infinite variety, we probably need a whole box of tools, not just a hammer and screwdriver.

A few further words are needed, however, to describe and define what I mean by "complex" development challenges, since of course doing almost *anything* in development is complex: building roads, irrigating fields and immunizing babies are all really hard things to do—by anyone anywhere. But truly "complex" problems go a step further than being complex in the technical or logistical sense, because roads, fields and babies don't vote, can't go on strike, can't be corrupted, can't change their minds and can't wage organized campaigns resisting (or supporting!) what is being done to them. Only people can do these things. Moreover, truly complex problems have people not only as the "objects" of change but the "subjects" by which change is realized: justice requires judges or juries to make discretionary decisions, often on the basis of deeply imperfect evidence (different people might decide differently); emergency health-care workers have to make literal life-and-death decisions about how to respond most effectively to victims of accidents or violence (mistakes can be fatal); to educate a child through high school takes approximately 12,000 hours of face-to-face interaction with people we call teachers, all of whom have to take general guidelines and requirements ("the curriculum") and decide how to optimally engage with dozens of students, all with different temperaments and learning styles. In such situations, it's often not at all obvious what the "right" response is; one just has to start by trying **s***omething*, and then work iteratively towards what becomes or emerges over time as the right response.

Evaluating interventions in this space is harder still. Truly complex interventions have no observable "counterfactual", so standard procedures for doing "rigorous" assessment are essentially impossible. Such interventions unfold over trajectories that are mostly likely highly variable (and non-linear) across time and space, making calls about their impact, in the absence of a defensible theory of change, conditional on the semi-arbitrary point in time at which the follow-up data is collected. These structural characteristics problematize not only claims to causality (internal validity), but broader concerns about generalizability and scaling-up (external validity)—and it is these latter concerns on which I wish to focus. Methodology per se, even (or especially) "rigorous" methodology, does not solve these problems as manifest in complex interventions. As Nancy Cartwright and Jeremy Hardie (2012: 137) astutely put it,

> *the bulk of the literature presently recommended for policy decisions… cannot be used to identify 'what works here'. And this is not because it may fail to deliver in some particular cases [; it] is not because its advice fails to deliver what it can be expected to deliver… The failing is rather that it is not designed to deliver the bulk of the* ***key facts*** *required to conclude that it will work here. [emphasis added]*

What are these "key facts" needed to discern whether a given intervention might work "here", and how might such facts be acquired? Let me suggest that there are three discrete realms of "key facts" evaluators need to acquire, and that these are optimally discerned by integrating evidence via an integrated array of methods.

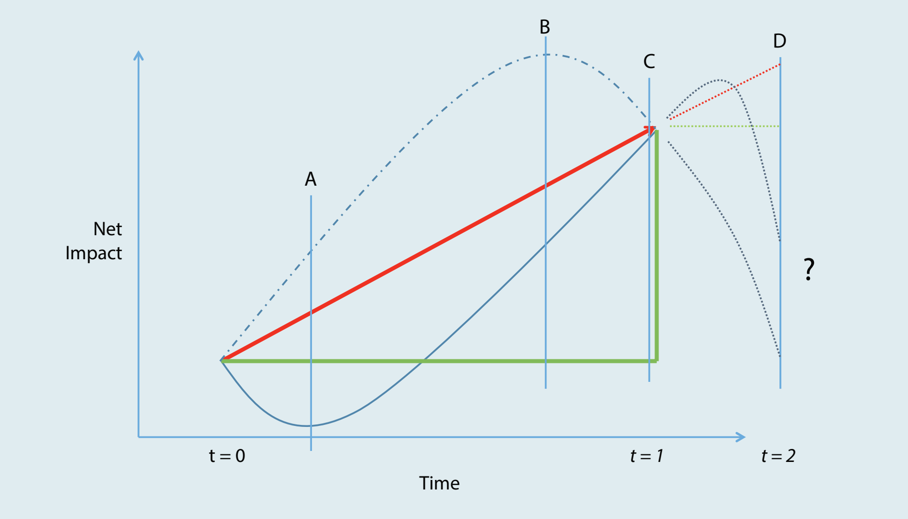
The first such fact, I suggest, is *implementation capability*—can the designated agency tasked with delivering the policy, programme or project actually do so? Even if impeccable evidence from elsewhere strongly suggests that, say, cash transfer programmes or micro-credit schemes have significantly reduced (say) poverty, and your government has decided to prioritize poverty reduction, the introduction of these "proven" interventions from afar are only going to be as good as their implementation. The content and design quality of programmes are obviously important, but these features per se are deeply insufficient for determining the outcomes as experienced by targeted groups. One might be slightly more confident that relatively "simple" interventions will be duly implemented, but the higher the level of complexity—as defined above—the harder (by definition) it will be for the designated agency to implement it, and thus the lower the likelihood that it will be uniformly well implemented at scale. Indeed, this argument, coupled with evidence from simulations (Eppstein et al 2012) and experiments (Pritchett and Sandefur 2015), essentially leads one to conclude that the external validity of complex interventions should be assumed to be zero.

Even so, the pragmatic reality is that policymakers and practitioners engage in external validity challenges all the time—compelling ideas and evidence addressing complex policy problems, no matter where they come from, must be taken seriously. In the face of this imperative, a second key fact for evaluators to consider is what I shall call *contextual compatibility*. That is, given sound design quality, adequate financial and political support and a capable implementation unit, the intervention itself must still enjoy local legitimacy: targeted groups in particular must deem the intervention to be consistent with their values, aspirations and concerns—or, more specifically, the intervention must be a coherent and credible response to a problem that targeted groups themselves have nominated and prioritized. It is for this reason that a given intervention's claim to being a global "best practice" becomes deeply problematic—if international development "experts" deploy such reasoning as warrant for introducing a particular intervention in response to a complex and contested development problem (e.g., by claiming that "rigorous evidence" elsewhere regarding the intervention's efficacy thus deems it a "best practice", and that skeptical or risk-averse policymakers should thus adopt), then in due course it is highly likely to be either rejected outright or rendered ineffective.[163] Much work is needed to discern that a proposed intervention is indeed contextually compatible.

The third domain of key facts evaluators of complex interventions need to be aware of is *reasoned expectations* regarding by when outcomes should be discernable. As noted above, complex interventions are highly likely to follow decidedly non-linear (even deeply idiosyncratic) trajectories as they unfold, meaning that, absent knowledge of where an intervention should be by when, claims about "impact" are going to be contingent on the semi-arbitrary point at which the evaluation is conducted. Per Figure 1 below, an evaluation team conducting its assessment on four different interventions at points "A" or "B" would reach four different conclusions about the effectiveness of each one, ranging from outstanding success to actively making things worse. Once one relaxes the assumption—which is otherwise ubiquitous in evaluations of development interventions—that the change trajectory is monotonically linear and increasing, then it should be apparent that almost *any* judgement call about efficacy (and thus generalizability) is dependent on engaging with reasoned expectations about what one would expect—on the basis of experience, evidence or theory—to have happened after a particular period of time.

In short, if your intervention (say, justice reform) entails high levels of discretion and face-to-face interaction, requires considerable implementation capability, has low contextual

FIGURE 1. THE IMPORTANCE OF UNDERSTANDING IMPACT TRAJECTORIES



[163] See Bridges and Woolcock (2017) on such outcomes in public sector reform in Malawi.

compatibility and unfolds along an uncertain trajectory, then making singular claims about impacts that are solely attributable to the intervention's design characteristics per se is deeply problematic, as is the capacity to generalize about the intervention's likely impact elsewhere, and/or at scale. *In this space, case studies and process tracing are essential tools for evaluators (or at least for key members of the evaluation team.*)

Let me conclude with several important implications and applications that I think follow from what I've argued here. First, evaluators (and researchers more generally) should take the analytics of external validity claims as seriously as we do internal validity. At present our profession functions at graduate-school level on the latter but at kindergarten level on the former; indeed, too often we (erroneously) presume that the "more rigorous" our identification claims, the stronger the warrant this provides for making claims about generalizing and scaling up. But that is just not so; even our identification strategies are suspect, it seems to me, if we have not adequately made impact claims conditional on knowledge (or reasoned expectations) of likely impact trajectories over time. Identification is just one issue among many needed for policy advice.

Second, evaluations need to expand the (vast) array of social science tools available for rigorously assessing complex interventions. Within *and beyond* economics, RCTs [randomized control trials] are just one tool among many. New literature on case studies (Gerring, Goertz), QCA [qualitative comparative analysis] (Ragin), complexity (Ramalingan, Kaufmann) and especially "realist evaluation" (Pawson, Tilly) need to be taken vastly more seriously than they are if we are to adequately engage with complex interventions. Third, all policy professionals need to figure out how to make implementation cool; it really matters—any intervention is only as good as its implementation. Learning from intra-project variation is key way in which this might be done; projects themselves should be seen as laboratories, as "policy experiments" (Rondinelli 1993). Evaluators also need to promote greater understanding of *how*, just not whether, interventions work—this will entail forging a stronger focus on mechanisms and theories of change. Fourth, no matter if the primary concern is internal or external validity, claims about the efficacy of complex interventions cannot be undertaken in the absence of what we might call a "counter-temporal" (not just counterfactual): that is, a reasoned sense of where we should expect a given intervention to be after a certain time period.

Fifth, and finally, no one in the business of assessing complex interventions can (or should want to) avoid the imperative to generalize and scale up (or not). We already have interesting documents with examples of local successes that failed when scaled up (business registration in Brazil), of mediocre local projects that, at scale, became a national flagship programme (community development in Indonesia), of projects that, *on average,* had little impact but, upon further interrogation, had positive effects for some groups and negative effects on others (livelihoods project in India). What we need to know in each of these instances is *why* and *how* such outcomes prevailed; deploying a mixed methods evaluation strategy in dialogue with social theory can provide fruitful avenues by which to find and share answers.

Thank you very much for the opportunity to be with you, and to share some thoughts on this important topic.

REFERENCES AND SUGGESTED READINGS

Andrews, Matt, Lant Pritchett and Michael Woolcock, *Building State Capability: Evidence, Analysis, Action*, New York, Oxford University Press, 2017.

Bamberger, Michael, Vijayendra Rao and Michael Woolcock, 'Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development', in Abbas Tashakkori and Charles Teddlie (eds.) *Handbook of Mixed Methods* (2nd revised edition) Thousand Oaks, CA, Sage Publications, 2010.

Bamberger, Michael, Jos Vaessen and Estelle Raimondo (eds.) *Dealing with Complexity in Development Evaluation*, Sage Publications, 2015.

Bamberger, Michael, Jim Rugh and Linda Mabry, *RealWorld Evaluation: Working Under Budget, Time, Data and Political Constraints (2nd ed.)*, Sage Publications, 2013.

Barron, Patrick, Rachael Diprose and Michael Woolcock, *Contesting Development: Participatory Projects and Local Conflict Dynamics in Indonesia,* New Haven, Yale University Press, 2011.

Bridges, Kate and Michael Woolcock, 'How (not) to fix problems that matter: assessing and responding to Malawi's history of institutional reform', Policy Research Working Paper No. 8289, Washington, DC, World Bank, 2017.

Brixi, Hana, Ellen Lust and Michael Woolcock, *Trust, Voice and Incentives: Learning from Local Success Stories in the Middle East and North Africa,* Washington, DC, World Bank, 2015.

Cartwright, Nancy, and Jeremy Hardie, *Evidence-Based Policy: A Practical Guide to Doing it Better,* New York, Oxford University Press, 2012.

Eppstein, Margaret L., Jeffrey D. Horbar, Jeffrey S. Buzas, and Stuart A. Kaufmann, 'Searching the clinical fitness landscape', *PLOS One*, 7(11): e49901, 2012.

Pritchett, Lant and Justin Sandefur, 'Learning from experiments when context matters', *American Economic Association: Papers and Proceedings* 105(5): 471-475, 2015.

Rao, Vijayendra, Kripa Ananthpur and Kabir Malik, 'The anatomy of failure: An ethnography of a randomized trial to deepen democracy in rural India', World Development 99(11): 481-497, 2017.

Woolcock, Michael, 'Toward a Plurality of Methods in Project Evaluation: A Contextualized Approach to Understanding Impact Trajectories and Efficacy', *Journal of Development Effectiveness* 1(1): 1-14, 2009.

Woolcock, Michael, 'Using Case Studies to Explore the External Validity of Complex Development Interventions', *Evaluation* 19(3): 229-248, 2013.

Woolcock, Michael, 'Reasons for Using Mixed Methods in the Evaluation of Complex Projects', in Michiru Nagatsu and Attilia Ruzzene (eds.) *Philosophy and Interdisciplinary Social Science: A Dialogue,* London, Bloomsbury Academic, forthcoming.