Dealing with Complexity in an Increasingly Interconnected World

Michael Woolcock

Development Research Group, World Bank

National Evaluation Capacities Conference

Istanbul

20 October 2017

Pop quiz! Putting evaluations to work…

1. In Bolivia, a carefully designed and implemented pilot intervention provides cash payments to poor rural families who send their 10-year-olds to school each day. A ‘rigorous’ evaluation finds a significant improvement in the children’s test scores. The education minister is keen to showcase a national ‘flagship’ initiative, and has resources available. Do you advise her to scale up the pilot?

Pop quiz! Putting evaluations to work…

1. In Bolivia, a carefully designed and implemented pilot intervention provides cash payments to poor rural families who send their 10-year-olds to school each day. A ‘rigorous’ evaluation finds a significant improvement in the children’s test scores. The education minister is keen to showcase a national ‘flagship’ initiative, and has resources available. Do you advise her to scale up the pilot?
2. A recent study published in a prestigious evaluation journal, using a large cross-country sample, finds that countries exiting from regional trade agreements significantly improve their ‘rule of law’ score in subsequent years. Fiji’s Minister of Justice desperately wants to improve his country’s global ranking on the rule of law index, to encourage foreign investment. Do you advise him to push for Fiji’s exit from its regional trade agreements?

Pop quiz! Putting evaluations to work…

1. In Bolivia, a carefully designed and implemented pilot intervention provides cash payments to poor rural families who send their 10-year-olds to school each day. A ‘rigorous’ evaluation finds a significant improvement in the children’s test scores. The education minister is keen to showcase a national ‘flagship’ initiative, and has resources available. Do you advise her to scale up the pilot?
2. A recent study published in a prestigious evaluation journal, using a large cross-country sample, finds that countries exiting from regional trade agreements significantly improve their ‘rule of law’ score in subsequent years. Fiji’s Minister of Justice desperately wants to improve his country’s global ranking on the rule of law index, to encourage foreign investment. Do you advise him to push for Fiji’s exit from its regional trade agreements?
3. A randomized control trial (RCT) of a large women’s empowerment project in Bihar (northern India) finds that, on average, the intervention had no effect after two years. Do you recommend shutting it down?

Pop quiz! Putting evaluations to work…

1. In Bolivia, a carefully designed and implemented pilot intervention provides cash payments to poor rural families who send their 10-year-olds to school each day. A 'rigorous' evaluation finds a significant improvement in the children's test scores. The education minister is keen to showcase a national 'flagship' initiative, and has resources available. Do you advise her to scale up the pilot?

   Small to Large

2. A recent study published in a prestigious journal, using a large cross-country sample, finds that countries exiting from regional trade agreements significantly improve their 'rule of law' score in subsequent years. Fiji's Minister of Justice desperately wants to improve his country's global ranking on the rule of law index, to encourage foreign investment. Do you advise him to push for Fiji's exit from its regional trade agreements?

   General to Specific ('There' to 'Here')

3. A randomized control trial (RCT) of a women's empowerment project in Bihar (northern India) finds that, on average, the intervention had no effect after two years. Do you recommend shutting it down?

   Interpreting Non-Impact

In each case, the right answer is: "It depends!" But on what, exactly?

What else do we need to know to provide better answers?

How/where can we find it?

Evidenced-based Policy

(as conventionally understood, at least by researchers…)

- **Development policy, practice plagued by…**
  - Inadequate, low-quality ‘hard data’
  - ‘Soft’ methodologies. Thus,
    - Too much reliance on ‘anecdotes’, ‘advocacy’
    - Lack of ‘rigorous evidence’ on ‘what works’.

  As a result,
    - Finite public resources deployed inefficiently
    - ‘Aid effectiveness’ debates fester, go unresolved
    - Taxpayers, politicians remain skeptical, cynical

  Thus we need a more ‘scientific’, ‘gold standard’ approach
    - “To do for development what RCTs did for medicine”
      - Elite researchers as “white lab coat guys” in development

Really?

Improving decision-making in development

"[T]he 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."*

> Nancy Cartwright and Jeremy Hardie (2012) *Evidence-Based Policy: A Practical Guide to Doing it Better* (New York: Oxford University Press, p. 137)

What 'key facts' do decision-makers need?

How might these 'facts' be acquired?

For what kinds of interventions are these 'facts' especially important?

Overview

- 1.4 cheers for ‘Evidence-based policy’
  - (As conventionally understood)
    - Its virtues, its severe limits
- Complexity when everything is complex
  - Defining characteristics of complex interventions
  - Assessing their internal validity (net impact)
  - Assessing their external validity (if it works there, then here?)

➢ *Different challenges need different types of evidence*

- Some examples
- The future of development is only more complex
  - In poor, middle-income and rich countries alike
  - Expanding the ecosystem of evaluation options

Evidence-based policy:
1.4 cheers (at best) for all that

- More, better data always desirable
- Sound methods always beat sloppy methods
- Accountability for use of public resources is vitally important
- Raising professional standards, meeting high expectations is a virtue
- ‘Evidence-based Policy’ well-suited to accurately assessing standardized interventions (e.g., traffic flows)
  - Such evidence can indeed yield ‘best practices’
  - Some such ‘best practices’ can be readily generalized, scaled

But also serious problems and limits

- Not how today's rich countries became rich
- Not how today's rapidly growing poor countries accelerated
- Often hugely expensive, time-consuming
  - Practitioners mostly need good-enough data now, not in three years
- Numerous ethical concerns, sometimes legitimate political resistance
- Potentially strong on answering 'whether' something works, on average…
- …but often weak on
  - 'How', 'why', 'for whom' an intervention works
  - Deciding between alternatives, optimizing under (many, vexing) constraints
  - Discerning "causes of effects" (cf. "effects of causes")
  - Building capability for policy implementation (cf. policy design)
  - Generalizing, scaling: Will it work here? Will bigger be better?

Often inadequate for assessing 'complex' interventions

Complexity when everything is complex…

- Most complex interventions (or elements) characterized by:
    1. **High discretion** (agency, choice)
    2. **Transaction-intensive** (many face-to-face interactions)
    3. **Impose obligations** (cf. deliver a service)
    4. **Unknown solutions** to prevailing problems

Which *inherently* yield highly variable outcomes:

- Over time, contexts, groups, implementing agencies
    - e.g., schooling, justice, empowerment, governance
- How to more adequately assess such interventions?

Matching *types* of evidence to *types* of problems

- Establishing causality is *really* hard, even in the actual world of “white lab coats”…
  - let alone let alone development interventions
  - let alone *complex* development interventions
- Problems can be usefully arrayed by the nature, extent of their ‘causal density’
  - the number of discretionary, human interactions involved
    - i.e., from particle physics (zero) to medicine (some) to families (numerous)

Consider physics…

“Only the first nine pages in the 33-page article, published on 14 May in *Physical Review Letters,* describe the research itself — including references. The other 24 pages list the [5154] authors and their institutions.
The article is the first joint paper from the two teams that operate ATLAS and CMS, two massive detectors at the Large Hadron Collider (LHC) at CERN, Europe’s particle-physics lab near Geneva, Switzerland. Each team is a sprawling collaboration involving researchers from dozens of institutions and countries.
By pooling their data, the two groups were able to obtain the most precise estimate yet of the mass of the Higgs boson — nailing it down to ±0.25%.”

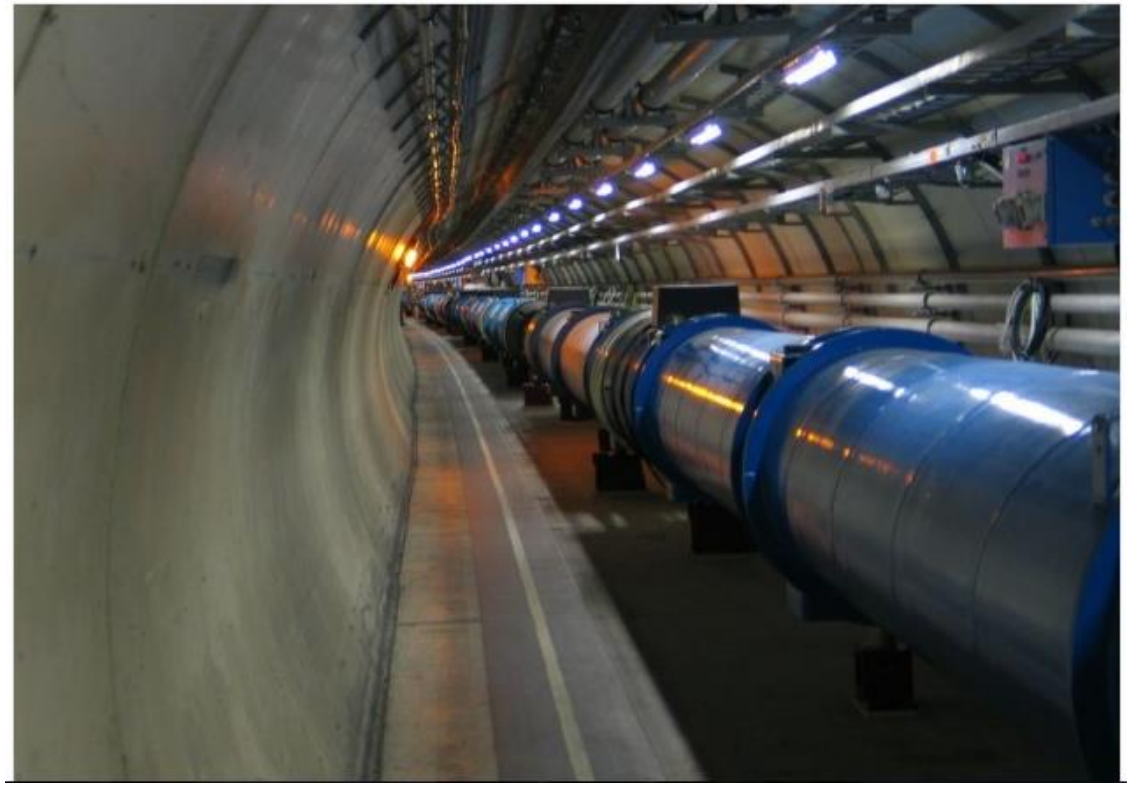
NATURE | NEWS

Physics paper sets record with more than 5,000 authors

Detector teams at the Large Hadron Collider collaborated for a more precise estimate of the size of the Higgs boson.

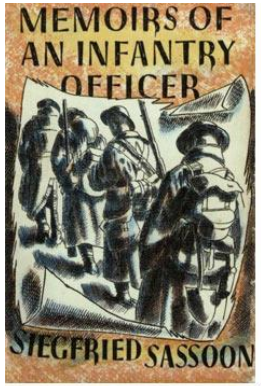
Davide Castelvecchi

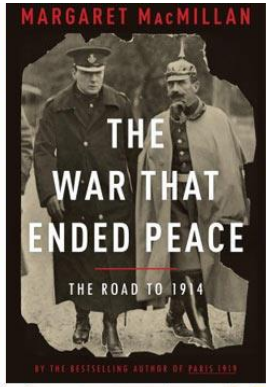
15 May 2015

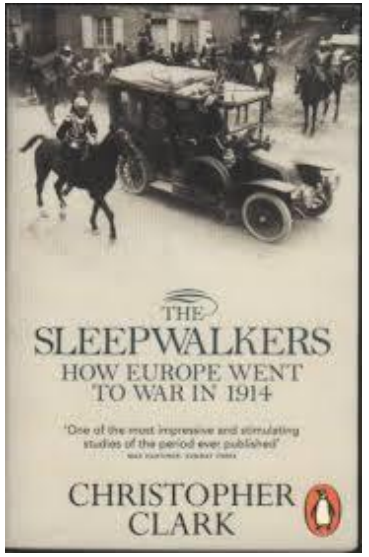


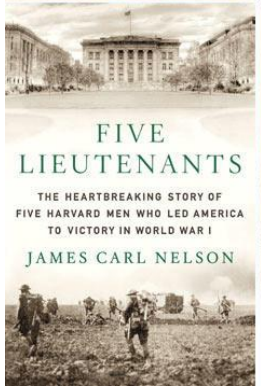
Consider historical questions. Historians ask: What 'caused'…

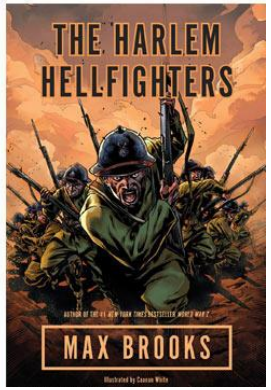
- The doctrine of the 'divine right of kings' to fall?
- The advent of civil, human and gender rights?
- The Industrial Revolution? World War I?
- The end of slavery? of colonialism? of apartheid?
- Independence of Haiti, India, South Sudan?

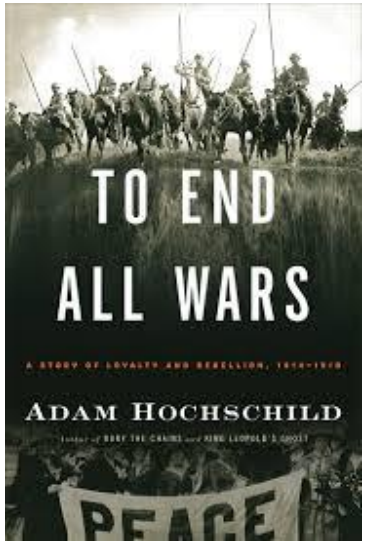






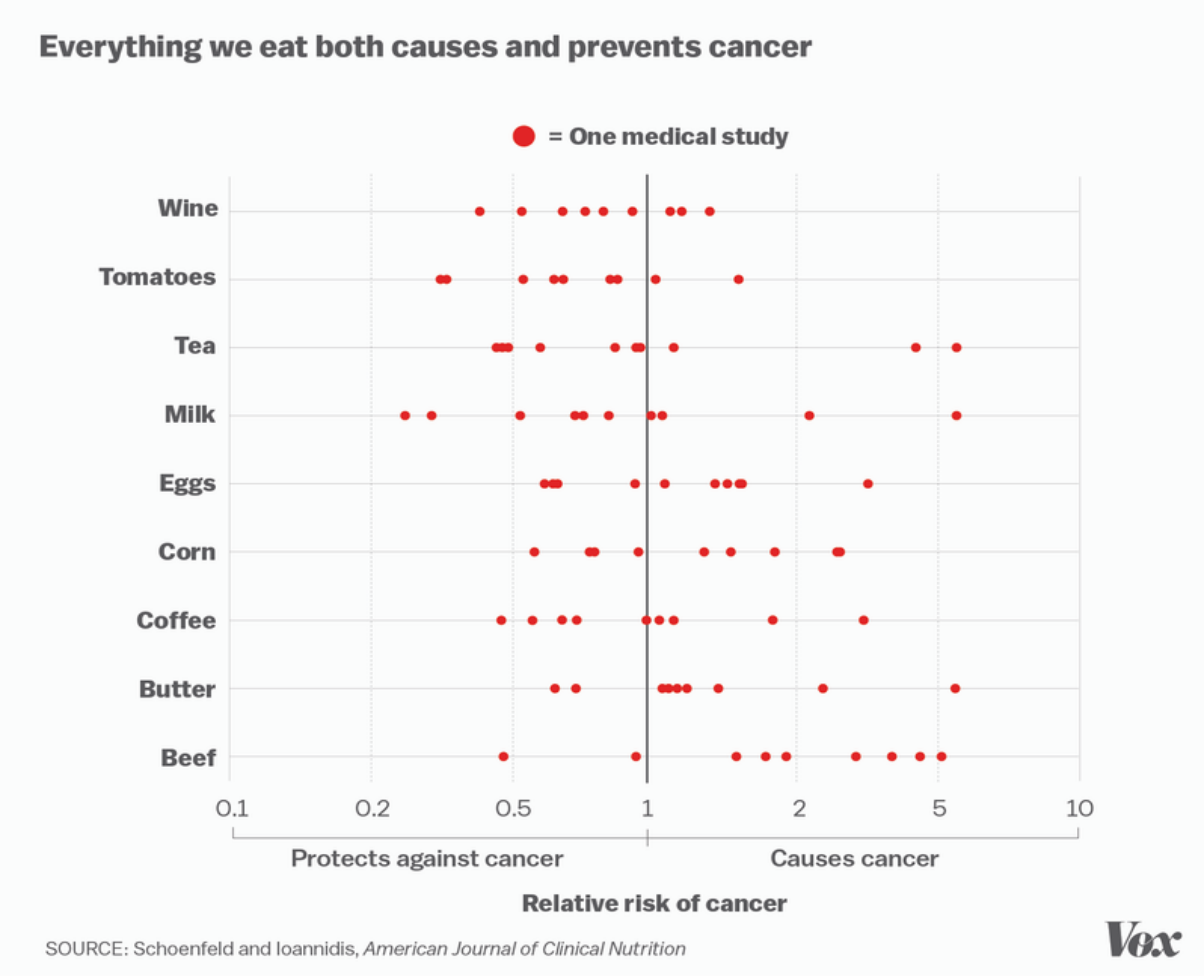






- Lots of things… so they tell a 'conjunctural causation' story
- So too for 'complex' development activities: careful *process tracing* and counterfactual reasoning can identify the sequence of causal mechanisms (inside 'the black box') connecting certain variables (and not others)
- **Absence of 'rigorous methodology' is NOT why these problems aren't 'solved'**

Consider causality, extrapolation in medicine (‘moderate’ causal density)



- Extrapolating from RCTs in drug trials
  - Rothwell (2005)
- **The case of ‘Black 6’**
  - Engber (2011), in *Slate*
    - Black 6 turns out to be “a teenaged, alcoholic couch potato with a weakened immune system, and he might be a little hard of hearing.”
  - Seok (2013), in *NYT*
    - “Years and billions of dollars” compromised

Which sports are assessed most 'rigorously'?

Key point: Problems determine methods, not the other way around

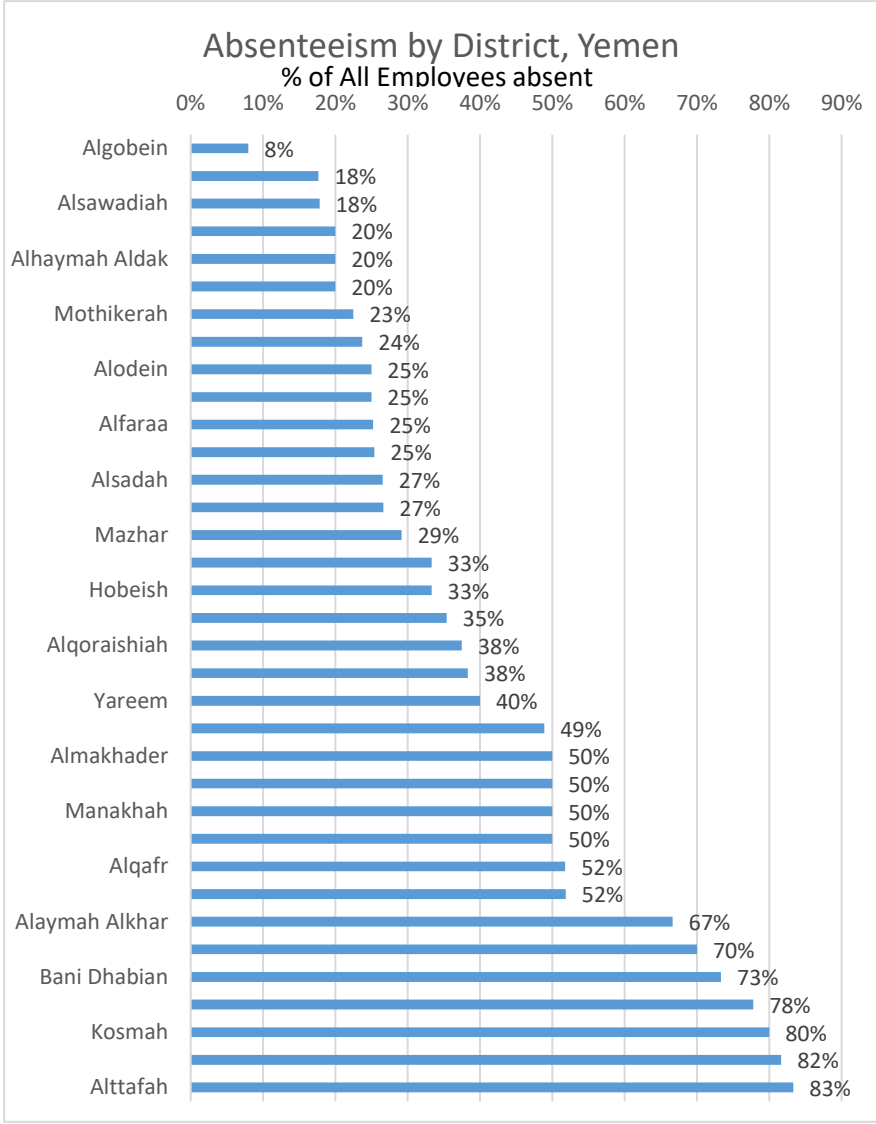
Making, extrapolating impact claims

Quality of empirical knowledge claims turns on…

1. Construct validity
   - Do key concepts (‘property rights’, ‘poverty’) mean the same thing to different people? What gets “lost in translation”?
2. **Internal validity…**
   - Controlling for other factors potentially shaping the result
   - E.g., Selection effects: Programs rarely placed randomly…
3. …assessed against **a ‘theory of change’**
   - How a project’s components (and their interaction) and processes generate outcomes
   - *Reasoned Expectations*: where by when?
4. **External validity** (how generalizable are the claims?)
   - If it works here, will it work there? If it works with this group, will it work with that group? Will bigger be better? (Does 10x get you 10y?)

The internal validity challenge in ‘complex’ projects

- Outcomes inherently dependent on implementation capability
- By design, interact with / respond to “context”
  - Huge unobserved heterogeneity
- Highly variable impact across time, space, groups
  - Even when carefully designed, faithfully implemented, adequately funded, politically supported
  - Can construct ‘mean’ impact, but more insightful is the ‘standard deviation’
- No true counterfactual, except other instances of themselves
- Orthodoxy struggles to explain success and (especially) failure
  - “Unhappy projects are unhappy in their own way”



What to do?

Endogenize research into implementation

- Experiment, learn, iterate, adapt
  - India's Social Observatory (Vijayendra Rao et al)
    - http://www.worldbank.org/en/programs/social-observatory
    - A social science of delivery
  - Global Delivery Initiative
    - http://www.globaldeliveryinitiative.org/

Discern the "causes of effects", not just "effects of causes"

- Understand *how, for whom* (not just whether) impact is achieved
  - e.g., Process Evaluations, or 'Realist Evaluations'
    - e.g., Pawson; Rogers; Barron et al
  - Knowledge claims require mixed methods, theory, and experience
    - RCTs can be usefully deployed to assess certain aspects
    - Complement to, not substitute for, orthodox approaches

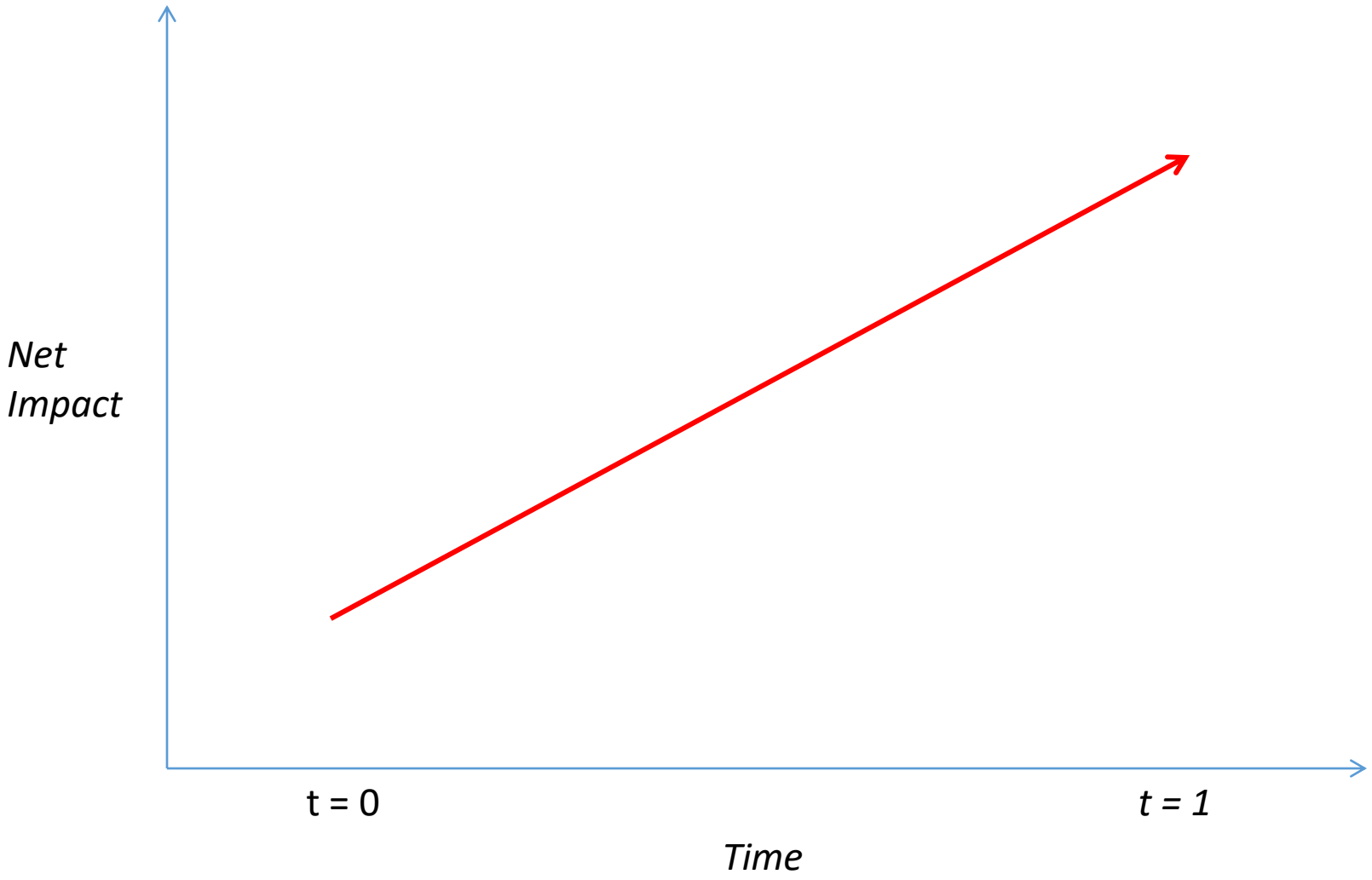
Crucially, also need a theory of change

When nothing seems to help, I go back and look at the stonecutter hammering away at his rock perhaps a hundred times without as much as a crack showing in it. Yet at the hundred and first blow it will split in two, and I know it was not that blow that did it — but all that had gone before.

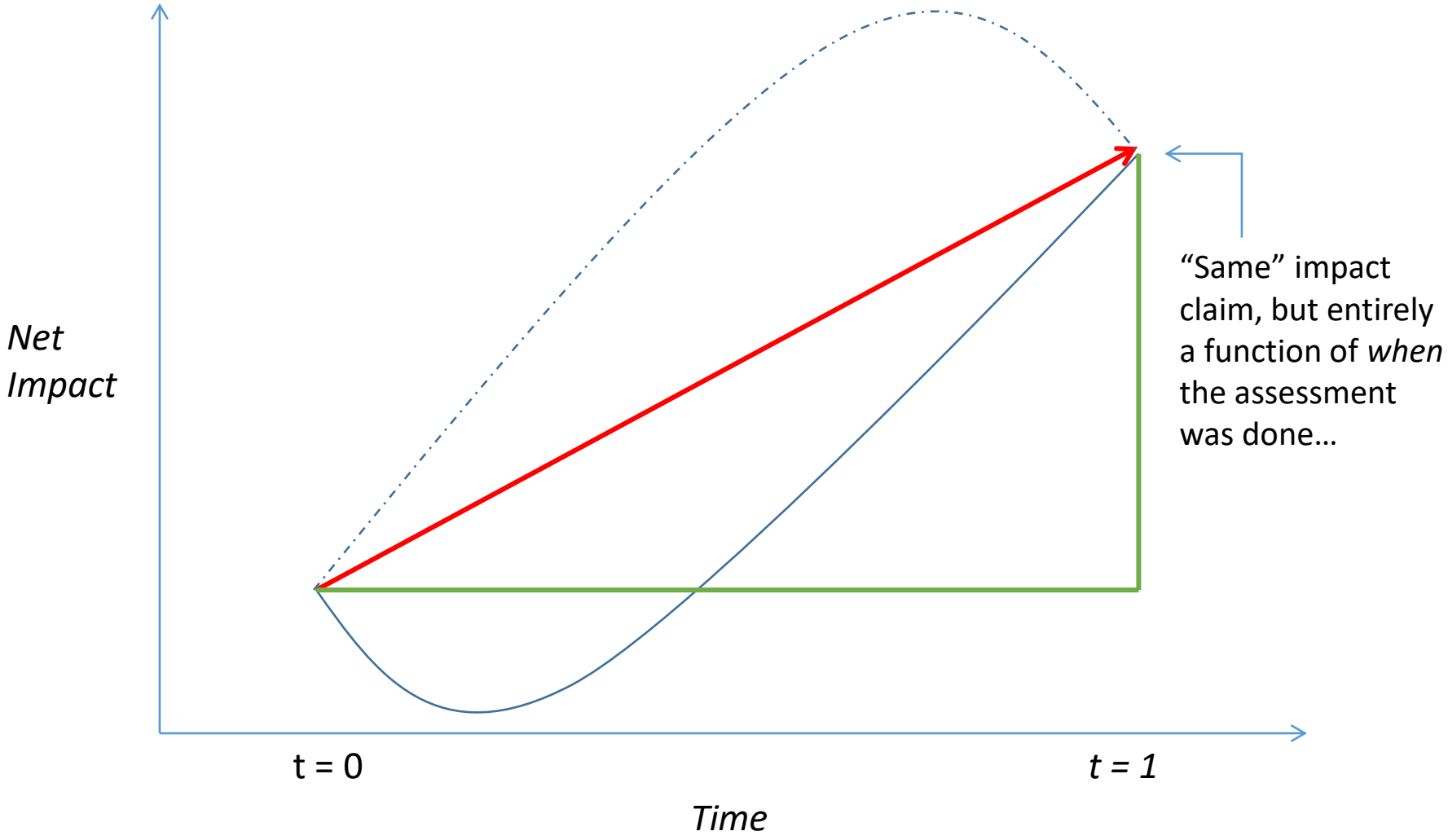
Jacob Riis

Sunflowers vs acorns

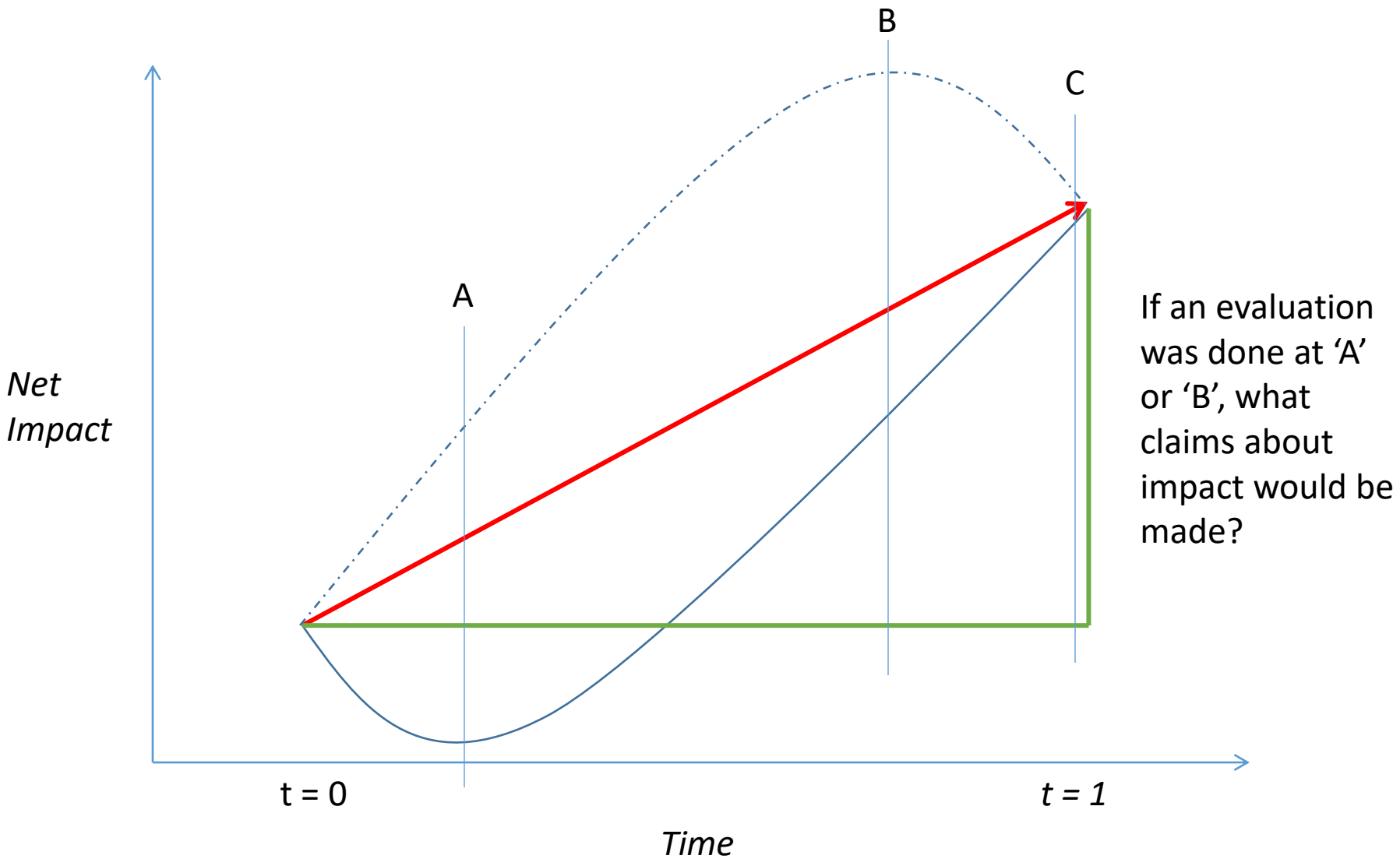
Understanding impact trajectories



Understanding impact trajectories

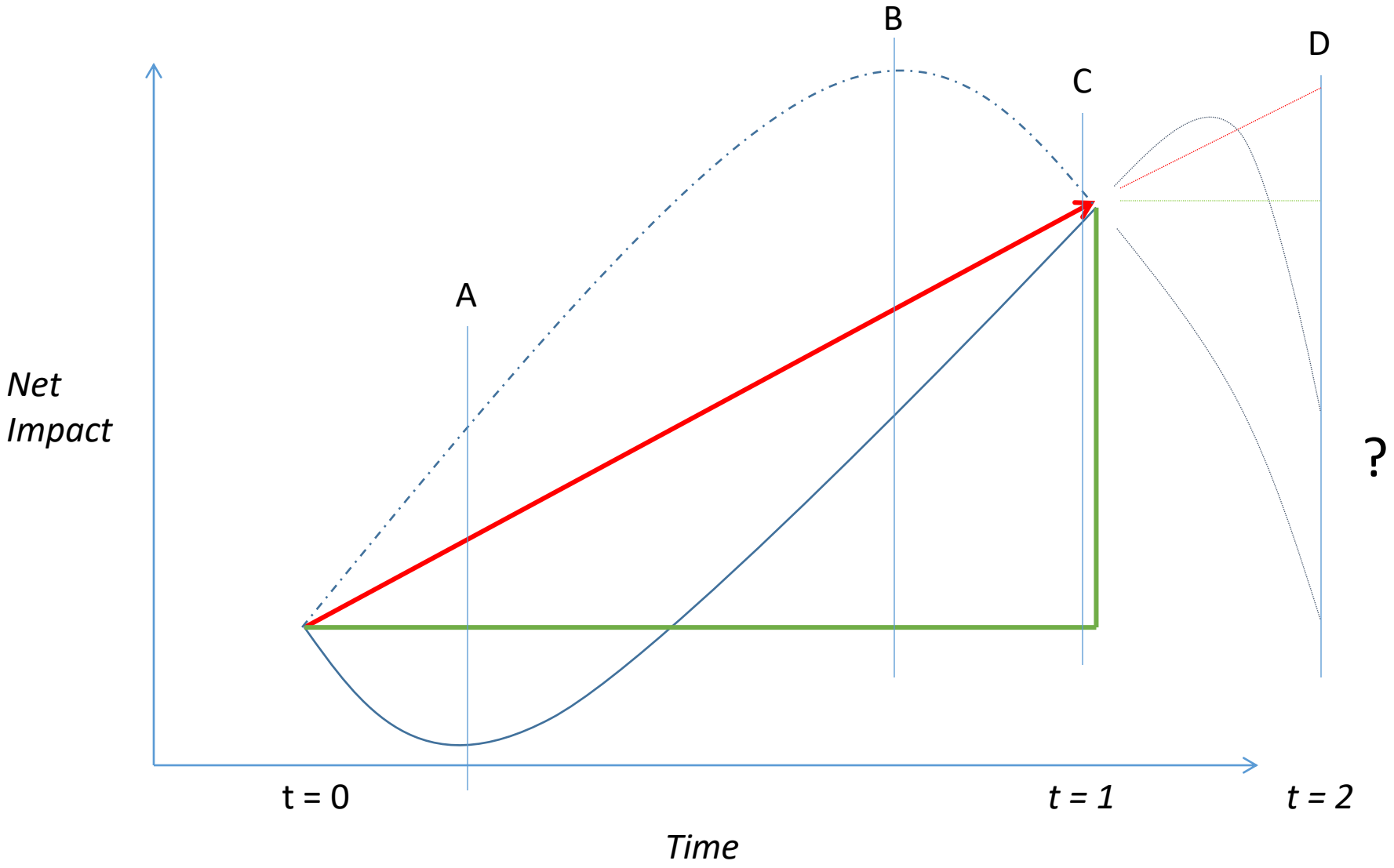


Understanding impact trajectories



If an evaluation was done at 'A' or 'B', what claims about impact would be made?

Understanding impact trajectories



External validity's "key facts"

1. **Causal density of intervention** (its type/level of 'complexity')
2. **Implementation capability** (can the designated agency do it?)
3. **Contextual compatibility** (local legitimacy)
4. **Reasoned expectations** (impact trajectory)

Bottom line: if your intervention (say, justice reform)…

has high causal density
requires high implementation capability
has low contextual compatibility, and
unfolds along an uncertain trajectory,

then assume generalizability is low (probably zero).

In this space, case studies and process tracing are essential tools.

Implications

- **Take the analytics of EV claims as seriously as we do IV**
  - Identification one issue among many needed for policy advice
- **Expand the (vast) array of social science tools available for rigorously assessing complex interventions**
  - Within *and beyond* economics
    - RCTs as one tool among many
    - New literature on case studies (Gerring,), QCA (Ragin), Complexity (Ramalingan)
  - See especially 'realist evaluation' (Pawson, Tilly)
- **Make implementation cool; it really matters…**
  - Learning from intra-project variation; projects themselves as laboratories, as "policy experiments" (Rondinelli 1993)
    - 'Science of delivery' must know *how*, not just whether, interventions work (mechanisms, theory of change)
    - Especially important for engaging with 'complex' interventions
      - Problem-Driven Iterative Adaptation (PDIA) (Andrews, Pritchett, Woolcock 2013)
- **Need 'counter-temporal' (not just counterfactual)**
  - Reasoned expectations about what and where to be, by when?

Two applications

- **Small to Big**
  - Local success that became a national failure (Brazil)
  - Mediocre local project that became a global flag-bearer (Indonesia)
  - Why?
- **Interpreting Non-Impact**
  - Livelihoods project (India), assessed by RCT, yielded no overall effect
  - Why?

Common Lesson: Explaining (the past) and advising (the future) about both the effectiveness of complex interventions requires knowledge of *key facts*:

* Design quality and characteristics
* Context idiosyncrasies
* Implementation capability

All of which, in turn, requires integration of mixed methods, theory, experience

Suggested reading

- Ananthpur, Kripa, Kabir Malik and Vijayendra Rao (2014) 'The anatomy of failure: an ethnography of a randomized trial to deepen democracy in rural India' World Bank Policy Research Working Paper No. 6958
- Bamberger, Michael, Vijayendra Rao and Michael Woolcock (2010) '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
- Bamberger, Michael, Jos Vaessen and Estelle Raimondo (eds.) (2015) *Dealing with Complexity in Development Evaluation* Sage Publications
- Bamberger, Michael, Jim Rugh and Linda Mabry (2013) *RealWorld Evaluation: Working Under Budget, Time, Data and Political Constraints (2nd ed.)* Sage Publications
- Barron, Patrick, Rachael Diprose and Michael Woolcock (2011) *Contesting Development: Participatory Projects and Local Conflict Dynamics in Indonesia* New Haven: Yale University Press
- Brixi, Hana, Ellen Lust and Michael Woolcock (2015) *Trust, Voice and Incentives: Learning from Local Success Stories in the Middle East and North Africa* Washington, DC: World Bank
- Woolcock, Michael (2009) '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
- _____ (2013) 'Using Case Studies to Explore the External Validity of Complex Development Interventions' *Evaluation* 19(3): 229-248
- _____ (forthcoming) '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