the LAB
MARKET SYSTEMS DEVELOPMENT FOR DECENT WORK

Fooled by randomisation: why RCTs might be the real 'gold standard' for private sector development

International Labour Organization

#innovation
#lablearning
#rcts

Market systems development for decent work - "the lab" - is an initiative of the International Labour Organization (ILO).

We test innovative ways to better measure and maximise pro-poor employment outcomes through a market systems development approach.

Read about us here:
www.ilo.org/thelab

Written July 2014, revised July 2015. The lab is funded by the Swiss State Secretariat for Economic Affairs (SECO) and run by the ILO's Small and Medium Enterprise Unit. The views expressed here do not necessarily represent those of SECO or the ILO.

You're probably one of the many who've been told that randomised control trials (RCTs) are the 'gold standard' methodology for assessing impact. Sadly, the people who say this to you do so without any sense of irony, or indeed history.

The original gold standard was a monetary system where currency was based on a fixed quantity of gold. For a long time considered imperfect and subject to constant debates about its effectiveness, the gold standard was eventually abandoned for a better system - the fiat system of banknotes in your wallet right now - in the early 20th century (Claasseen, 2005). Not a single country in the world uses it anymore.

RCTs may be much more like this gold standard than their champions would like to believe, and end up being consigned to the history bin in a similar manner. Critiques of randomisation are not new, and have been well-documented[1]. They are expensive, deliver big data dumps rather than more rapid feedback, require a consistent and homogenous treatment, don't say much about why change happens, and lack context. Let's even leave aside the ethical issues for now.

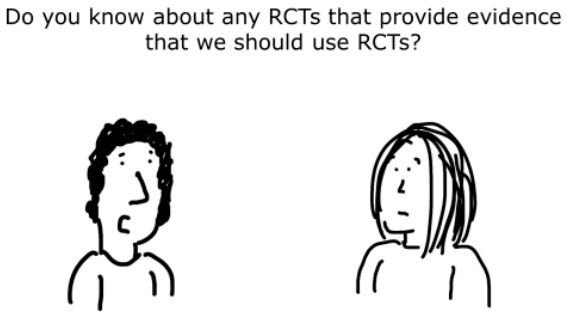
This, however, has not stopped a huge rise in the number of experimental studies being commissioned. A 'randomista' movement, which came out of medicine and into social science, has largely focused on the development impact of health and microfinance interventions but is now creeping into the field of enterprise development. But before we get preoccupied with whether or not RCTs could fit into this field, it's worthwhile stopping to think whether or not they should.

These scientific experiments have the potential to do as much harm as good, especially if they're cookie-cuttered into a private sector development (PSD) context. Here's why.

Use? We'll worry about that later

First, there is no evidence that evidence from a randomised experiment carries any more weight in policy-making. The UK Department for International Development's chief economist, for example, says their policy decisions are made on the basis of a compelling case based primarily in theory, since any evidence-base, no matter how rigorous, will always be incomplete (Johnson, 2013). A lot of money is currently being spent on RCTs for little added value in terms of influence. In words that randomistas will relate to: the additionality of these trials is dubious.

[1] See, for example, the work of Angus Deaton at Princeton University



freshspectrum.com

This is borne out by the historical relationship between RCTs and social policy. According to Jones (2010), experimental research was fashionable in the USA in the 1960s until it was clear that it couldn't deliver all that was hoped. The world gave up on these designs, like it gave up on the monetary gold standard, a long while ago. Pritchett (2014) remarks that "strangely, whether or not decades of social policy RCTs actually did have impact on policies and outcomes in the USA just kind of never came up in arguing that they would in developing countries". Few donor countries currently use RCTs in their own social policies - so it seems bizarre that these are now being foisted on the 'recipients' of their aid.

The use of RCTs in informing development practice fairs little better. Ramalingam (2011) finds that "despite the volumes of impact evaluations [in the last decade], much remains unchanged in the aid sector…mistakes are still seldom acknowledged and frequently repeated…research is still frequently side-lined, bad news is still buried, and the lack of results is still not publicised". One reason for this might be a significant publication bias: Jones (2010) claims "over 95% of published RCTs show 'positive' impact, which severely limits the ability to really learn". Better methods are clearly not a magic fix for the inability of the development community to actually use and take up knowledge.

In the words of Ramalingam (2011), it is clear that "evidence – regardless of its origins – is not systematically absorbed into policy and practice". The problems are deep-rooted. The 'build it and they will come' mentality has not worked in terms of improving learning, and no amount of taxpayer money pumped into re-creating 'laboratory' conditions for scientific experiments will change that (Ramalingam, 2011).

If all you have is a hammer

Randomista researchers have a vested (and understandable) interest in promoting experimental designs: it's how they make their money, and forge their reputations through peer-reviewed publications. Development agencies, however, have fallen hook, line and sinker for the pitch, loving the idea of finally having a way to quantify their impact, and jumping on the RCT bandwagon because it seems to confer 'credibility' (Jones, 2009).

This has often been to the exasperation of practitioners in the field. For starters, projects need improved systems and resources to do better real-time measurement, allowing them to be able to understand why things are (or are not) working out *as they go along* in order to react and adjust accordingly. It doesn't take a multi-year trial to hear back from training recipients that the curriculum was inappropriate, or that there is no demand for what they are being trained in. At the moment, even basic monitoring is so poor that jumping straight to RCTs is like going from driving a battered Ford Fiesta to a brand new Ferrari.

Sadly, RCTs often eat up the budget and appetite for continuous learning, driving a deeper wedge between data and action rather than bringing them closer together. Randomisation also demands the roll-out of a fixed and pre-packaged treatment (a product or service), limiting the ability of projects to change interventions as they go along. This jars with recent trends in

hikingartist.com

development and beyond to adopt a more flexible approach to help navigate uncertain operational environments: through tighter feedback loops to build-measure-learn (Reis, 2011);, to fail-faster but learn-faster in adaptive management (Harford, 2012); or to rapidly test new ideas in Problem-Driven Iterative Adaptation (Andrews, Pritchett and Woolcock, 2012).

RCTs are most suited to 'traditional' output-driven programmes with ambitions of national scale, like distributing insecticide-treated nets to reduce malaria. Inherently 'complex' areas - including governance, capacity building, market development, policy influencing - are not well-suited. Yet even here there is a push to shoehorn these into RCT designs, choosing narrow elements of programmes that are randomisable and honing in on them for impact measurement. At best this takes resources away from getting a more holistic understanding of the overall impact of a programme, and at worst risks artificial design and 'locked in' implementation just to satisfy a methodology.

Working out how to fit RCTs into private sector development is an upside-down way of looking at things. The question should instead be: what methods are best suited to the context and research needs of the project and intervention? Research questions must drive research methods, not the other way around (USAID, 2010).

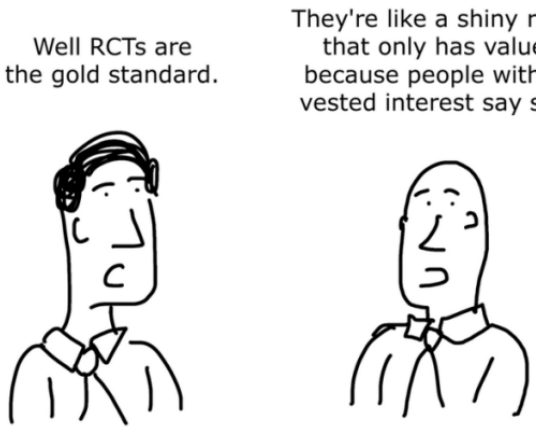
After all, says Jones (2010), "RCTs tend to be carried out where they are methodologically convenient rather than where the new knowledge is really needed". If all you have is a hammer, everything will start to look like a nail. Overreliance on a small cadre of research organisations specialising in only one form of research design, no matter how much of a 'buzz' they are currently creating, will condemn us to ignorance in the name of rigour (USAID, 2010).

Considerable knowledge does need to be generated on the efficacy of private sector development approaches: not just for 'policy' purposes, but so that projects themselves can react and adapt – in real-time as far as possible – in response to the actual impact they are having. This is especially the case for market-sensitive and demand-led projects that seek to build incrementally towards change, and shy away from the temptation to roll out one-size-fits-all development 'fixes' which try and impose pre-set solutions (but which, conversely, may be easiest to 'randomise'). To paraphrase the former USAID Administrator Andrew Natsios, be careful not to confuse the ease of measurability with development significance (Ramalingam, 2013).

Rigour is not random

Those wielding the RCT hammer often claim that randomised designs are top-of-the-pile (Patton, 2011). Setting the standard of rigour so high, randomistas see everything else as an inferior second tier and game for methodological criticism (Johnson, 2013).

But according to the UK's Overseas Development Institute, the knowledge that results from one impact evaluation methodology is just as applicable and potentially useful as the knowledge from any other kind of methodology (Ramalingam, 2011). All methodologies are equal, and none are more equal



freshspectrum.com

than others. Voices that advocate only quantitative, experimental independent evaluation as the only defensive means of assessing impact are simply wrong (Taylor, 2013).

Rigour - the quality of being extremely thorough and careful - is not a binary concept or the domain of one particular methodology. It is in fact a matter of degree. As recognised by USAID (2010), there is more than one acceptable way to look at impact, and insightful and credible results can be obtained by assessing projects using a variety of methods providing different aspects of rigour.

It is time to "reclaim rigour" for more qualitative and other mixed methods that are better suited to environments where most PSD interventions take place (Green, 2013). Researchers, practitioners and policy-makers alike need to get much better at counting what counts, not just what can be counted (Green, 2013). There is much work to be done. Promising theory-based methods, for example, ironically remain much stronger in theory than they do in practice. But a critical first step is realising that there is value in investing in this broader range of methods. After all, says Taylor (2013), "that something is difficult to measure should be motivation to find an alternative approach rather than to abandon it all together" or to fall lazily back on the latest methodology fad.

This is why in the ILO lab we're currently working with projects to test out alternative ways of impact measurement, including theory-based designs (using methods such as Mayne's contribution analysis), quasi-experimental designs (using difference-in-difference techniques) and more rigorous impact monitoring (using the DCED Standard for Results Measurement)[2]. We are also innovating around ways to better track impact through systemic change – which explicitly aims at catalysing the kind of 'spill-overs' and 'contamination' that gives randomista researchers such headaches.

RCTs have their place in the toolkit. Projects looking to roll-out standardised, targeted and controllable 'treatments' - like a set training package - may be feasible for randomisation. Whether these trials will offer valuable insight will depend on what kind of knowledge is being sought under what certain circumstances. It is clear, however, that RCTs are just one tool among many. Do not make the mistake of equating rigour with randomisation.

Learning how to learn

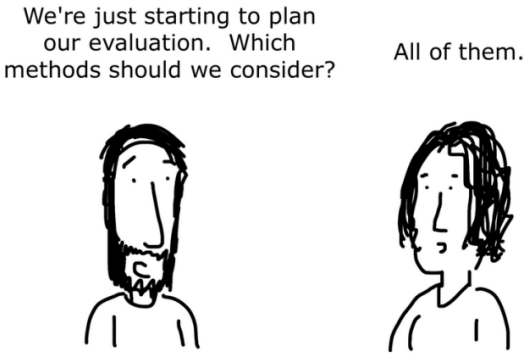
According to Claassen (2005), when medical scientists adopted the term 'gold standard' to describe their tests, they confused the meaning. Inspired by the Olympics, where the best athlete won gold, they stuck 'gold' to 'standard' (meaning an authoritative or recognized exemplar of quality) to denote something that was the best in the world. But this absolute meaning detracted from the original monetary meaning, which was a relative measure of the best available under certain conditions. As used today, the term is ambiguous to the point of being redundant.

[2] For other methods that are appropriate to estimate attribution in complex environments, see White and Phillips (2012)

The only 'standard' that does exist is one of methodological appropriateness: putting the function of what we want the research to find out before the form of what methods we will use to do so. Initiatives such USAID's degrees of evidence framework have begun to make headway in this, helping users to understand the myriad of methodological choices that are 'on the table' and the trade-offs these choices require. All social science research, after all, takes place outside of the sterile conditions of a laboratory and involves trade-offs based on political, resource and contextual conditions.

Regardless of which methodology is deemed most appropriate, the root cause of the current inability in international development to learn has relatively little to do with the availability or quality of evidence or the level of rigour. It is likely much more about the system of incentives and structures underpinning how development agencies and donors operate, which together create the conditions for the use - or more accurately, the misuse or complete lack of use - of impact knowledge.

According to Ramalingam (2011), the "key factors in impact utilisation are…human, organisational and political" as much as they are purely technical. In this sense, we all need to be more willing to *learn how to learn* (Ramalingam, 2011). Until we do this, many rainforests will be decimated to churn out a conveyor belt of impact evaluations, whether randomised or not, that will not be used seriously to improve development policy and practice.



freshspectrum.com

References

Andrews, Matt; Pritchett, Lant; and Woolcock, Michael. June 2012. "Escaping Capability Traps through Problem-Driven Iterative Adaptation".

Claassen, JAHR. 2005. "The gold standard: not a golden standard" (British Medical Journal, 2005 May 14; 330(7500): 1121)

Green, Duncan. 14 May 2013. "Redesigning Aid for Complex Systems". Blog available at: http://oxfamblogs.org/fp2p/how-to-plan-when-you-dont-know-what-is-going-to-happen-redesigning-aid-for-complex-systems/

Harford, Tim. 2012. "Adapt: Why Success Always Starts with Failure".

Johnson, Susan. 2013. "Randomistas and microcredit: shutting the evidence gate after the policy horse has bolted". Blog entry at:

https://cdsblogs.wordpress.com/2013/01/28/randomistas-and-microcredit-shutting-the-evidence-gate-after-the-policy-horse-has-bolted/

Jones, Harry. 2009. "The 'gold standard' is not a silver bullet for evaluation". Available at: http://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/3695.pdf

Jones, Harry. 2010. Comment on a blog post by Green, Duncan. 7 May 2010. "Randomised Control Trials: Panacea or Mirage". Available at: http://oxfamblogs.org/fp2p/randomized-controlled-trials-panacea-or-mirage/

Patton, Michael Quinn. 2011. "Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use"

Ramalingam, Ben. 2011. "Learning how to learn: eight lessons for impact evaluations that make a difference". Available at: http://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/7096.pdf

Ramalingam, Ben. 2013. "Aid on the Edge of Chaos"

Reis, Eric. 2011. "The Learn Start-Up: How Constant Innovation Creates Radically Successful Businesses"

Taleb, Nassim Nicholas. 2005. "Fooled by Randomness: The Hidden Role of Change in Life and Markets"

Taylor, Ben. 2013. "Evidence-Based Policy and Systemic Change: Conflicting Trends?". Available at: http://www.springfieldcentre.com/wp-content/uploads/2013/07/Evidence-Based-Policy-and-Systemic-Change1.pdf

Pritchett, Lant. 2014. "An Homage to the Randomistas on the Occasion of the J-PAL 10th Anniversary". Blog available at: http://www.cgdev.org/blog/homage-randomistas-occasion-j-pal-10th-anniversary-development-faith-based-activity

USAID. 2010. "Time to Learn: an evaluation strategy for revitalized foreign assistance". Available at: http://pdf.usaid.gov/pdf_docs/Pnadw234.pdf

White, Howard and Phillips, Daniel. June 2012. "Addressing attribution of cause and effect in small n evaluation frameworks: towards an integrated framework". Available at: http://www.3ieimpact.org/media/filer_public/2012/06/29/working_paper_15.pdf