Bridging to Action Requires Mixed Methods, Not Only Randomised Control Trials

DOI:
10.1057/s41287-019-00201-x

Document Version
Accepted author manuscript

Citation for published version (APA):
Olsen, W. (2019). Bridging to Action Requires Mixed Methods, Not Only Randomised Control Trials. *European Journal of Development Research*, *31*(2), 139-162. https://doi.org/10.1057/s41287-019-00201-x

Published in:
European Journal of Development Research

Bridging to Action Requires Mixed Methods, Not Only Randomised Control Trials

Wendy Olsen January 2019

European Journal of Development Research – accepted for publication. The estimated publication date is April 2019.

Abstract:

Development evaluation refers to evaluating projects and programmes in development contexts. Some evaluations are too narrow. Narrow within-discipline impact evaluations are weaker than multidisciplinary, mixed methods evaluations. A two-step process leads toward profoundly better arguments in assessing the impact of a development intervention. The first step is setting out the arena for discussion, including what the various entities are in the social, political, cultural and natural environment surrounding the chosen problem. The second step is that once this arena has been declared, the project and triangulation data can be brought to bear upon logical arguments with clear, transparent reasoning leading to a set of conclusions. In this second step we do need scientific methods such as peer review, data and so on. But crucially, the impact evaluation process must not rest upon a single data type, such as survey data. It is dangerous and undesirable to have the entire validity of the conclusions resting upon randomised control trials, or even a mixture of data types. Different contributions to knowledge exist within the evaluation process, including the interaction of people during action research, ethnography, case-study methods, process tracing and qualitative methods. The cement holding my argument together is that multiple logics are used (retroductive, deductive, and inductive in particular). Deductive mathematics should not dominate the evaluation of an intervention, as randomised controlled trials on their own lend themselves to worrying fallacies about causality. I show this using Boolean fuzzy set logic. An indicator of high quality development evaluation is the use of multiple logics in a transparent way.

Key words: randomised control trials, comparative case-study research, methodology, retroduction, mixed-methods impact evaluation

Author's contact: wendy.olsen@manchester.ac.uk

Title: Bridging to Action Requires Mixed Methods, Not Randomised Control Trials

Introduction

The reliance on Randomised Control Trials (RCT) for the impact evaluation of development projects is growing, but their cost levels exceed what is required, and the alternative methods are routinely underfunded. Our aim is to bridge from problem to data to action (Morgan and Olsen, 2007 and 2008). Bridging to action means making sure that evidence gathered can enable warranted arguments to emerge from the evaluation process. That process must have elements of action, participation and monitoring, as well as data gathering. If there are control groups, and there need not be, development practitioners may not be aware of the control group contrastive findings until the very latest stages of the project when the action research and survey research elements are brought together. Yet important points about how to intervene, what works where, and who is being affected, should be brought to the eyes of the development organisation as soon as possible – not after a long delay. Thus, using RCTs alone is often going to be unethical and too slow, as well as being unwise.

To put the debate into perspective, I have broken down the impact evaluation process into stages (Figures 1 and 3). These are the stages that the scientific dialogue usually takes. When we are being honest and transparent, the scientific method itself generates arguments that may take the same structure.

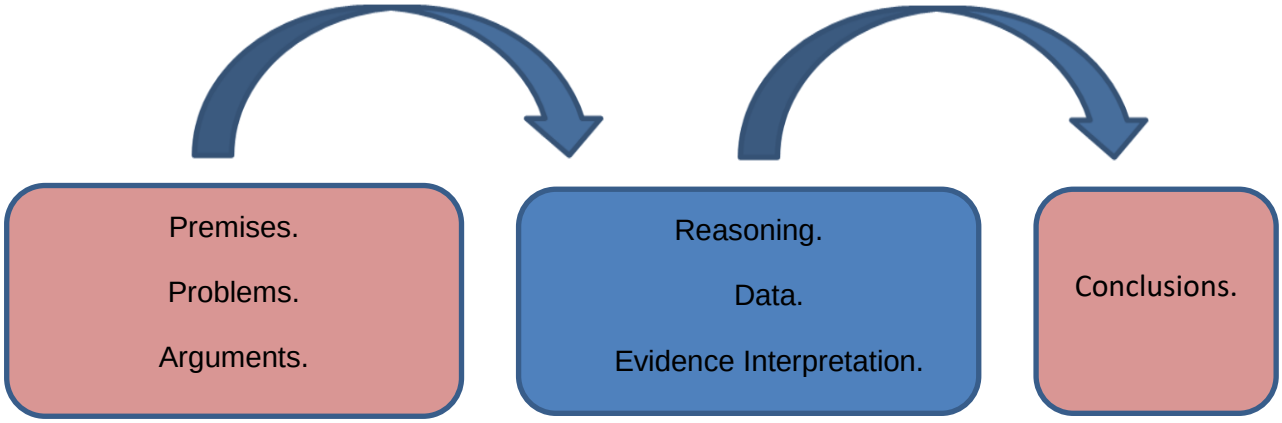

Figure 1: The Evaluation Stages 1 and 2: The Arena Stage, Data Stage, and Conclusion

PPA is the arena and RDEI is the debating stage. Yet even at the stage of identifying the problem to be solved by the intervention, arguments will occur. Opinions will differ about how best to proceed. Two disciplines may disagree on what this is all about. Two sets of practitioners may disagree on the best means of changing development practice. But we agree to disagree, having settled the terms of our debate. This early stage includes ontological exploration: what exists, what pre-exists the present, what is considered to be relevant amongst all the extant entities, and therefore what areas of life matter. (This will then imply which disciplines will be brought to bear.)

In this paper I argue that the second and concluding stages involve collective reflection and reflexivity, so we actually return to Stage 1to reconsider the framing and then generate even more data: data that is both new and different. Therefore Figure 3 is a better representation (having feedback loops) than Figure 1.

Evaluation Stage 1

In the arena stage, we are reading the literature, talking about our knowledge, and discerning what are the objects, histories, entities, narratives and meanings that we are going to investigate. We find out the black lines that distinguish difference in the world. The entities to which we refer are real – examples such as caste, class, gender are all controversial, but the controversies are about the world, not just a bunch of voices arguing in a void. The “ontic” is what exists. The project’s ontology is its sense of where it is, what cases are involved, who are the actors, what processes they start off, what histories matter.

Next, the impact evaluation project goes on to do some activities, and people learn from these, and generate both data and reasoning about the findings. The stage of data interpretation is not only about interpreting quantitative data; it is about learning (Befani, Barnett, and Stern, 2014). Crucially it is also a debate about what words to use, what concepts to apply, what are true narratives and what are misleading ones.

A secondary stage, third stage, and fourth stage, iteratively ask why these narratives, why these errors, how do we resolve disagreements? (Figure 3).

An example of how an evaluation agenda can be 'open' is found in Brunie, et al., 2014:

> Conduct a deep formative assessment to understand the needs, challenges, and values of the multiple potential programme stakeholders including direct and indirect beneficiaries. (Brunie, et al., 2014: 119)

The meaning of drawing a 'conclusion' is that the concluding claims rest firmly upon a mixture of all the preceding reasoning and data. If we take the advice of practical field researchers, such as Brunie et al., we will "Engage local partners for multi-sector interventions" (*ibid*.), and through discussions, reach agreement on what is ready for change.

If so, then multiple conclusions are both possible and feasible. There is more than one warranted argument, based on the same original arena of debate. Now we can have a scientific debate involving closer recourse to evidence. This does not mean depersonalising the debate. Some evidence will be heavily and intrinsically personal (as I will explain in the abduction section).

But for a sophisticated study worthy of respect, the 'arena' stage must be done carefully – integrating a review of literature – and the analysis stage must involve transparent evidence (Byrne and Ragin, eds., 2009). This has been recognised through the development of alternatives to RCTs: systematic reviews, meta-analysis and other initiatives such as "What Works ...", which offer useful tools to learn from previous experience.

Rebutting Three Misunderstandings

Some researchers have doubts about realism because it may seem to be too much a move from epistemological to ontological foundations. One thing that is often misunderstood about 'the real' and realism is whether or not the entities you look at, like fishing or Lake Victoria, are changing or fixed over time. Of course they are changing. Nevertheless, it is worth naming them. To name them is to select what is important from the background mass. A growing literature offers realist impact evaluation (Befani et al., 2014; Allmark and K. Machaczek, 2018; Olsen, 2009). The background has been set out explicitly by Maxwell and Mittapalli (2010) and implicitly by the case-comparative school (Rihoux and Grimm, eds., 2006; RIhoux, 2006; Ragin, 2008; and Snow and Cress, 2000).

Another misunderstanding occurs if one thinks realism is foundationalist, or that it favours the material world over the social world of social constructions. Far from it. (Maxwell and Mittapalli (2010) expand on this topic.) Social constructions are real in

the specific sense that the group labels found in a society have causal influence. There are causal mechanisms embedded in them. For example, the phrase *'ethnic groups of fisheries workers*', represents both the fisheries workers' ethnic identity as a cognitive self-reflection,and the social grouping that calls their group by an ethnic name as well as the real patterns of behaviour and culture behind ethnic differences. Therefore both labels and causes may be relevant. It's not foundationalist to be realist in our approach, but it does imply that we might at some point examine the evidence (by which I mean records of experience) to discern what these things are. In summary, although realism appears foundationalist, what is found to be 'real' is contingent on many contextual factors.

As an example,in the excellent analysis of the livelihood impacts of cash transfers across sub-Saharan Africa, Fisher et al. (2017: 306) pays attention to "Identity/status and inclusion/exclusion from [one's] network", by which they mean people's awareness of assigned ethnic and tribal groups, and social exclusion. These are not just social constructions. The society and its social institutions have real effects upon people, and these effects were brought to attention during focus groups.

In epistemological terms, just because the intervention evaluation is realist, it does not mean it is naively so. It does not mean we ignore social constructions, labels, naming conventions, translations or discourse. As Maxwell and Mittapalli explain, we use a scientific, not an "objectivist", realism. There is a sophisticated form of realism (often called scientific realism or transcendental realism) and there are forms of naïve realism (details found in Layder, 1993; Olsen, 2012; Olsen, forthcoming). The world should be seen dynamically.

Murshed-e-Jahan, et al. (2018) illustrate realism in a sophisticated value-chain study of fisheries in Nepal and Bangladesh. The realism in a study can be implicit, yet very helpful; and social constructionism of a weak kind can help in the participatory process by giving us a high awareness of meanings and discourses, not just what is factually asserted. By advocating participatory research, these authors advocate conversations.

The setting of an arena is also not a fixed, once-for-all stage. We can revisit Stage 1 after starting Stage 2; see Figure 3. This is commonly recommended in qualitative textbooks (Blaikie, 1993 and 2000). The arena lets us talk about our agenda.

Eight illustrations show how flexible iterative mixed methods have been proposed to be used in development intervention processes.

- Tremblay and Gutberlet (2010) show how community outreach, capacity building, and interviews helped in tandem over time to improve a recycling project.
- Luo, L. P., & Liu, L. (2014) argue that cultural awareness and touching base with local cultural differentiation are extremely important, and that contextually

well-grounded participatory research will help make the interventions effective.

- Nathan, S., Kemp, L., Bunde-Birouste, A., MacKenzie, J., Evers, C., & Shwe, T. A. (2013) used a "Research Reference Group comprising representatives from grant partners and participating schools formally met annually to advise on proposed study measures and recruitment of study participants." (2013: 3).
- Taylor, A., et al. (2012) showed that investing in leadership capabilities among water agency champions is a transferable, feedback-based method of invoking improvements, which can be transferred to improve other development change processes.
- Brink, M., et al. (2011) show how sustainable game management is not being achieved by results-oriented research, whereas process oriented, learning-based, adaptive co-management and co-regulation can work much better. They argue that not only is the new method better, but the old top-down methods are failing utterly to keep game reserves sustainable in either human livelihood terms or in the sense of sustaining health natural populations.
- Pollard, S. and D. DuToit (2011) argue that a practice-based understanding of policy, governance-sensitive actions, self-organisation, and feedback loops helped interventions in an integrated water resource management context to be much better (more effective, more acceptable, more responsive) over time.
- Gimenez, A. and A. Perez-Foguet (2010) use participatory methods to good effect in a policy impact assessment.
- Ngwenya, Barbara, Ntombi Ngwenya, Ketlhatlogile Keta Mosepele and Lapologang Magole (2012) show that short-term intervention studies miss out the key gender issues, including gross invisibility of women's work in a fisheries context in Botswana. Solutions are found through discussion.

Other concrete studies which use mixed data types effectively include Kambala, *et al.,* (2017), and King and Samii (2014).

In particular with reference to Ngwenya, et al. (2012), if we compare the randomised control trials in Mali by Masset and Gelli (2013), the methodological contrasts are immense. Ngwenya et al (2012) write in a transdisciplinary way with sensitivity and awareness of multiple stakeholder voices, whereas Masset and Gelli are aiming at a mono-disciplinary medical audience. The practice of publishing the RCT trial protocol first and the results later engages peer review in the Masset and Gelli (2013) case (see also Gelli, et al., 2017, 2018), but it does not promote any feedback loops and longer-term engagement of grassroots standpoint holders. It would be possible to do both, but apparently rigid epistemological boundaries, often known as 'epistemes' or sets of mutually exclusive rules about data, block the RCT users from invoking the better development research practices (Ravallion, 2009).

In brief, the three misunderstandings which often arise are: 1 determinism; 2 foundationalist naivety; and 3 closed stages.

The Evaluation Stage 2

Most good impact evaluations are going to use mixed methods, and gather mixed forms of evidence. In development contexts teams are guided in party by funding agencies such as Department for International Development (DFID) (see UK Aid Connect, 2018) to formulate a “theory of change", and use this to derive the strategy for causal analysis (and RCT design). The theory of change approach involves a discussion around how narrow/wide the net is to be cast for a project. Discussions which introduce theories of change include Funnell and Rogers, 2011; the Aspen Institute (2004); UNDP/Hivos (2011), cited in UK Aid Connect (2018).

I am often asked whether therefore the impact evaluations use pragmatism. Pragmatism was a key argument found in Creswell’s discussions of mixed methods overall (1994). Detailed summaries and critiques by Maxwell and Mittapalli (2010) and Allmark and Machaczek (2018) are helpful. The pragmatist ideas offered by Creswell and Plano Clark (2018: 38-40) are rather confusing. Creswell and Plano Clark cite Teddlie and Tashakkori (2003) in favour of abandoning metaphysical concepts, whereas the latter actually argued mainly in favour of abandoning the qual-quant paradigm wars (Teddlie and Tashakkori 2009: 8).

Overall I am not convinced of pragmatism making a strong contribution to impact evaluation. Pragmatism in philosophy means a number of things, and it offers very few implications for what you should do or may not do.

Instead when one argues for scientific realism making reference to the real world, this is a fierce and firm philosophical standpoint. Fierce because it demands that evidence be more than personal, that it be worthy of respect and scrutiny, and firm because it is foundational to consider that the world around us pre-dates us to any extent. It’s a metaphysical claim, whereas pragmatism makes few metaphysical claims. It is mainly about processes of making choices.

Implications of the Dialectic as Real

I would go even further. In general, the causality we observe in development does not work simply. It is complex for two reasons. First life moves onward through various dialectical processes, and secondly we ourselves as researchers are inside the changing world. We know that tensions exist, and we are able to influence the world, so these various tensions build up to social change. A dialectical change is one which has three stages – the initial stage, the building up of tension, and the resolution through some qualitative change. Often a project’s quantitative data avoids mentioning the very changes which are the ones that really shift a social situation.

Personal experience in social movements and political participation has proven to me that dialectical changes are a real, ontic thing, not just a figment of an author's imagination. Also, a multiple set of dialectics are going on all at once, making life challenging for all of us. I would argue that it is better to avoid being a determinist when it comes to causation. If this is the case, we then realise that evidence may belie the truth, or create a mask. For example, social class dynamics occur, so tension exists, so working class people may hide things from an elite observer. Language dynamics exist with a lot of tension during our modern period of rapid change, so we hide some facts that are best expressed in the minority language, and thus transcripts are faulty… and so on. We need to generate a critical perspective, enunciate our questioning views, check on all evidence, and get external reviews which may offer worthwhile insights. The core agents in an intervention are human, multiple agent interests. These combined to create a panoply. (The agents' voices do not just reflect arbitrary, subjective viewpoints; they offer glimpses of actual standpoints reflecting real interests.)

In turn, agents' values and beliefs affect what is accepted as premises. Or as data. Or as reasoning. To accept that we will study a problem, or an intervention, does not mean to accept agents' values and beliefs uncritically. It means to consider them critically. For a realist, the 'science' part involves comparing and making judgments about different arguments.

Interim findings from action researchers or participatory action research are of value in themselves. The findings from such activities usually are based on a mixture of retroductive and abductive logic. Retroduction involves asking why: why this activity choice; why this outcome occurred; and more deeply, why is it this problem that you feel needs to be solved (Downward and Mearman, 2007)? Abduction on the other hand refers to knowing 'from inside', in either a phenomenological or ethnographic context, what things really mean to people within the scene, such as a development project. Both these methods make use of personalised data. Instead of transparent or recorded data, both action research and ethnography depend heavily on a person's bodily experience, memories, and *ex post* vocalisation. Such methods must and can be combined with survey or interviews which involve much more record-keeping, allow comparability and nevertheless do not require a randomised form of control group selection.

The evaluation process overall does need to provide a transparent evidence base. Transparency arises naturally from survey data collection, re-use of schools' or NGOs' data, or the transcription of interviews or focus group transcripts. Meanwhile, at the concluding stage, the actors who are involved in the action research, monitoring & evaluation (M&E) or participatory research, must also re-evaluate their own positions (reflexively), and this is likely to occur in a private pre-publication dialogue, not In public. Admitting this will strengthen development policy debates,

not weaken them. Therefore, we could decide to use rapporteur methods to bring notes of the late-stage discussions into a public, ongoing reflexive multilogue.

Looking at the process I have just described, we use more than mixed methods. We use a mixture of steps of analysis interspersed with scene-setting decisions, re-analysis, induction from larger data sets in the survey components as well as the interview components, and deduction from the elements of quantitative data that belong in the overall database. Qualitative software, such as NVIVO, can be used to gather up these many threads. The human interactions of action research can be part of the ongoing evaluation efforts, and all documented into the NVIVO database for further retroductive analysis at the end.

To be really specific, the retroductive questions we use, in asking 'why' in a backward looking way, would include:

- Why did that not work?
- Why did this disagreement happen?
- What narrative reconstruction worked to resolve disagreements?
- Why were forecasts not met from early in the project? Was it due to some internal discursive limitation, or a clash with an external body? And, if so, what were the boundaries that had to be breached?

Comparative case analysis may fit well with the analysis of the survey data (Aus, 2009;). This need not be an RCT-based analysis. It could be a qualitative comparative analysis (QCA, fsQCA, csQCA) or a process tracing through both survey and other enquiry forms (Hellstrom, 2001).

All in all, my argument is that we must not limit one study to one logic from among the four choices:

{abductive, retroductive, inductive, deductive}.

We need to combine them and use them in sequence, or iteratively, as fits each study.

Furthermore, this sequence first has to have an arena for the discussion, and that is the most important initial part of an evaluation project.

Further Development of Retroduction

There are many ways to do retroduction. To retroduce being to ask why, we can first discern open retroduction and closed retroduction.

If you consider Stage 1 closed, and Stage 2 underway, you might only refer to existing recorded data (evidence) for retroduction. You may re-examine the existing variables, look for supporting quotes, find answers to obvious questions in the data

you have. This can include transcripts of interview data. This would be closed retroduction.

If you consider Stage 1 (arena–setting) to be open, then you can do open retroduction. You start to ask what needs to be re-conceived in order to get the answers to knotty problems in the research. Why did this intervention fail in this area? Why was it not implemented properly? What unexpected barriers were hit? As any experienced researcher knows, these questions involve re-opening the whole can of worms. This will mean a widened ontology; an openness to reconsider changing things that may have been considered fixed or irrelevant at the start. This would be open retroduction.

Important Background

RCTs are dominant, and examples abound. Costliness arises from the likely co-correlation of 'unobserved' confounders. The idea of a confounder implies closed retroduction. The costliness of the whole trial actually needs to be put on one side if we can do open retroduction to find out what is going wrong, or what went right, as quickly as possibly in a definitive way.

Cluster trial methods are widely used among those who believe in closed retroduction, no retroduction, or pure deductivism in research (Kelcey, Shen and Spybrook, 2016, and Taft, et al., 2012 illustrate these). The cluster idea locates groups of treated and untreated people (or cases, e.g. schools) far apart in space. Then no changes of strategy are allowed, as it would pollute the data. Diffusion, on the other hand, is a natural part of human communicativeness. Diffusion of ideas from the trial is discouraged by the RCT attitude, the RCT-committed team, the RCT protocol. See Kikuchi et al. (2015) for an example. This purity depends on a deductive logic. It assumes the 'data' will lie in a hermetic seal. Validity in this logic is not the same as validity in the transparent science sense. Validity in this logic is achieved by "if … data and trial then… therefore…" logic. But transparent scientific logic could take a different form: "The problem is XX and we discovered a barrier to solving it, BB…" There can be evidence about BB and XX, but it would not have been foreseen at the start.

My advice is to start Stage 2 but then allow for a revisit to arena-setting if need be.

This does not happen in cases like the RCT of Lubinga *et al*. (2014) or Pradhan *et al*. (2013).

Furthermore, basic reasons for the costliness of RCT trials is that the measurements must be both longitudinal (pre- and post-treatment measures, typically) and spread out far and wide to disallow people in the Control arm from discussing the treatment or its effects with those in the Treatment arm (an example in microfinance is McHugh, Biosca and Donaldson, 2017; see Orr, 2015). In other words, high-quality data is the key aim. This implies an epistemological value over other human values.

It turns out to be deontological. (Deontology refers to principles-focused thought.) I have no commitment to deontology, because it tends to be highly conceptual and not realist.

Secondary to this aim of pure data, the choice of 'cases' often turns into a reductive search for atomistic units of society to be 'affected' by the treatment. Atomism itself has been roundly critiqued in the philosophy of social science (Sayer, 2000). Whilst individualism is common in some circles of the academic world, known as having Anglo-Saxon traditions, most of the world's development science aims for a mixture of holism and atomism. Researchers try to achieve a healthy mixture of depth ontology and multiple levels so the atomic units are seen in context. A depth ontology is explained with useful diagrams by Lopez and Scott (2000: 33, 78, 88). In essence, we should assume open, interacting systems. I will give two examples illustrating the two extremes, and their costs.

On the atomistic side, examples like White (2013) give 'guidance' but implicitly send a message that holism is not wanted. On the mixed-methods side we have authors like Befani, et al. (2014) and Ssengooba, McPake and Palmer (2012) who argue in favour of more open forms of evaluation.

Alternative approaches also abound, notably participatory action research, action research, process tracing, qualitative comparative analysis (QCA), and monitoring and evaluation. Mixed methods can also combine these.

Ontological Discussion Should Occur Explicitly

Figure 2: Development Impact Evaluations Use Qualitative Methods Too Infrequently!

Source: Derived from a Web of Science literature search on development impact evaluation methods, 2018.

Creating an arena for debate is widely thought to be an epistemological activity. It's like trying to decide what we are talking about, so it seems it must be about knowledge. But really, constructing the conceptual 'ground' or fundamental list of entities and the key problem, and agreeing to talk about them for a while, is what I mean by setting up the arena. This is an ontological task.

Only once the ontology has begun to be worked out, with some boundaries on time and space, some aims and objectives, some concepts and names of things (such as who is a respondent, who is a participant, what members will be consulted, how action plans are to be written down if at all, and what will constitute evidence), then the project can really start in earnest as teamwork. Before this moment, there is speculation, there are beliefs, there are lay narratives. These are not to be thought of dismissively. But the nature of development as a science, a social science, is that it can focus on specific areas of life and bring together a discussion around the evidence and experience in these areas. Yes, that can include history. No, it cannot avoid setting up an 'arena' of debate.

Evidence is created as part of a bridge to action. Therefore, the control group evidence could be interesting, even if it were polluted by knowledge of the intervention, as long as it still creates contrasts that are of value in relation to objects or entities that we have agreed to notice.

It's important to realise that without action research or participation, people studying interventions are going to learn too slowly. The modes of learning must ideally be embedded in all the 'development methods' training. The tasks of research can be embedded in all the stages of the implementation of the intervention. Thus experts can work in the field alongside others, and everyone's expertise can be respected.

Otherwise, if we accepted the kind of standards of methods that are used by medics or by statisticians, who are focused on a set of identical cases reflected in the dataset [not the underlying reality], we could miss opportunities and waste a lot of research money. It is important how one argues about this.

I will now explain why I promote the use of process tracing, case comparison, context-focused Qualitative Comparative Analysis, and the discernment of causal mechanisms (see Sayer 2000 who sets out causality as real mechanisms).

These qualitative methods are consistent with choosing control groups with a restricted treatment group. What is not consistent is the use of merely deductive reasoning, or simplified purely mathematical methods, to draw policy conclusions.

Pro-Mixed Methods is Inherently Anti-Deductivist

Any sketch of the research process will tend to show mixed methods using multiple logics.

{abductive, retroductive, inductive, deductive}.

Induction in particular when combined with dialogue of diverse actors creates a dialogic element. This combination brings complexity to bear during the reflexive stages. The actors in the process bridge to action. I cited earlier examples from African fisheries and land-use planning, and sustainable development, where processes of change were mediated by change-agents whose growing awareness of multiple stakeholder standpoints and positions helped the group to move to better management practices. Bridging to action is a way to see development as practice. RCTs, on the other hand, postpone development practice and isolate the researchers from those who are being researched.[i]

Studying interventions may not be as important as getting to grips with empowerment issues, barriers to human capabilities, and development objectives which are fundamentally not being addressed by current research relationships. Research is human, not just medical. Development research about an intervention can be valid, highly structured, systematic, transparent and expert, without having a control group at all.

Thinking of Ngwenye *et al*. (2012), who recommended co-managing research whilst innovatively co-managing the fisheries resources, that study did have a large structured part. Ngwenye *et al.* used both a local survey and crosstabulations, and they triangulated their data with government data to give a backbone and generality to the study (2012: 111-116). The use of tabulated information is called systematic analysis (Olsen, 2012). At the same time, the researchers reflected on their interviews to reach new opinions, which they did not hold at the start. To argue that women are excluded from recognised roles in fisheries, and that new policies had obliterated women's roles, is to raise scientific, well-grounded objections.

In such cases of pluralist mixed methods, it is helpful to think of the author's logic not as deduction, which is arguing from a series of particulars to a general law, but rather a building up of complex arguments. 1. Many fishers are women, notably basket fishers (Ngwenya et al., 2012: 115). 2. Past policy on fishing favours men. 3. Favouring men occludes favouring women. 4. Not noticing women's work leads to the ignoring and discouraging of women's work. 5. Fish policy has discouraged women's work in fishing. 6. Favouring men will not help the fishers. The argument is logical but not deductive. This is known as a warranted argument: it moves from

premises to conclusions, with integration around key concepts. (A growing set of works by Alec Fisher, Bowell and Kemp, and Weston build up such logic into 'critical thinking', a skill which can be taught.)

In warranted arguments, the conclusion does not stand alone and it is not a subjective belief. Instead it rests upon the roots which are the premises of the argument.

It is possible to include treatment and control groups without limiting the research to RCT methods of statistical analysis. Agarwal (2018) illustrates by studying a sample of 70 groups of women, each doing rental farming collectively, in each of the states of Telangana and Kerala. She did follow-up interviews and had close liaison with the NGO and farmer leaders. She engaged in group-comparative statistical post hoc analyses of the overall outturn at the end of the study year. Here, there was no randomisation, but perfectly adequate comparative methods (*ibid*.) Agarwal's team also liaised with government officials and women's group leaders and members. This was teamwork with a scientist able to move about and conduct a cross-state comparative element.

Group reflexivity is generally important at the end of the control group data cleaning stage whether the evaluation uses case-comparative, group-comparative, or RCT methods. Here we are making systematic comparisons. However, conclusions from data tables alone are not final. The numbers cannot speak. An argument is something humans construct: it is not something an artificial intelligence could construct. Instead a circuit of discussion and revision is likely to be needed. We bridge to action again.

The ongoing nature of many interventions has meant a long delay in publishing findings of evaluations. This may be a good thing. Long long-term effects are not the same as short term effects, as Lam and Ostrom (2010) illustrate. Their study of watershed management in Nepal used data which showed short term water gain improvements below the water management engineering systems, but the longterm picture was the important scene really. Their paper concluded the longterm water provision was a key outcome.

Another reason for waiting to publish is that the overall pattern of economic or measurable effects may not be visible or evident to participants, but the project that has a wide scope and ambitious coverage requiring the assembly of a large and extensive dataset will have complex data. Once the project has happened, the team need to keep funding aside for their last few meet-ups, and translation is also needed so that outsiders do not dominate at this key stage. Pattern discernment often takes three formats at this stage:

1. Tables of means and comparison of means across groups.
2. Adjusted means comparisons, after allowing for confounders or while using IV adjustment or a fixed effects method. Sometimes a whole regression is

fundamentally based on a contrast of means by groups using 'difference in difference' (DID) methods. A DID estimate allows for change over time that is normal, versus a different trend among the treated group. A simple DID is a pair of lines moving upward showing a growth curve of profits over time (or scores or size), with the treatment group rising more quickly from the same base as the control group. Variants use curves, multiple treatment amounts, and so on.

Entropy based propensity score matching (PSM) is also used to make fair contrasts of group means, but PSM sometimes allows cases to drop out (see discussions in King and Nielsen, forthcoming; and Duvendack*, et al.,*2012). Textbook treatment of this topic by Hansen et al. (2011) and Lan and Yin (2017) offer innovations but not methodological pluralism.

3. Regression in full format, attributing causality to one factor or another. Here there is shifting ground for comparison over time. The initial random sampling is not enough to guarantee a base of support for claims focused on the treatment, because the base of support (ie the treated cases and their counterparts in the control group) may change due to compositional change over time. There are some problems with the regression format, too: do you use, or avoid, interaction terms? Do you allow moderation or not, in other words? Do you allow for mediation of effects? Isn't the treatment going to have multiple effects, and could any of these be external and not measured? If so, who records this, who speaks this truth, who brings it all together? Cress and Snow (2000) showed with sound ethnographic evidence that multiple outcomes may be meaningfully teased out, yet they would be ignored using the sophisticated atomistic statistical methods. Their study discovered 4 variants of policy success, R1 R2 R3 and R4, and then pursued the causal patterns for each for these. Such depth is too rarely achieved.

The truth is that adjustments to improve the accuracy of steps 2 and 3 have been promoted to the extreme of now questioning all tables of means. The very obvious descriptive point, that one group did better than another, which can be augmented by subtracting off the normal starting point, has been lost, and the audience has grown narrower and narrower.

Instead of merely 'using' 'control variables', we should consider the whole of how the treatment might take its effect. Changes in concomitant input variables would be expected. This is the 'coincident necessary part of a sufficient pathway' approach. Sufficiency of A for an outcome Y is analysed using a Boolean logic of factors that co-exist (A&B, also written A$\cap$B or A intersect B) and factors that are complements A$\cup$B, A or B. We write A=> Y if A is a sufficient condition for Y to have occurred. (An alternative relationship is A IFF Y, where IFF means if and only if, which is a stronger relation.) Most statistics assumes that if X IFF Y then Y IFF X, but I will show that this is a false and misleading assumption.

Once the data are ready (which is a late stage in a project) the great achievement is possible. The masterful statistician discerns a small rise in a good outcome, or a small decrease in a bad outcome, which wouldn't be noticeable to the naked eye or to participants, or without the regression controls or the adjustments.

There is now perhaps an 'evidence capital', a social capital of holding the evidence while deciding on the key findings, then noticing the key themes, and making sure they can be evidenced. This is like looking for something that might be obvious, like the emperor's new clothes.

It is deductivist to think we can't see the difference of outcomes using raw data (see Mock, *et al.,* 1993, pushing for case-control methods).

It's like saying we can only see Z in the pattern of X and Y if we use these cleaning methods, of which only an elite group will know.

Mixed Methods Authors Work in Teams

The numerous studies that follow DFID guidelines add a monitoring and evaluation aspect to each development project. DFID has invested in large numbers of evaluation experts. The idea of monitoring is not to make quantitative records, but to engage with participants and see how things are going at a midpoint and near the endpoint of an initial trials stage. Really listening is very important; it implies openness and narrative breadth, open questions, multidisciplinarity. These are all excellent bases for mixed methods but do require different skills from the experts in questionnaires and interviews. Therefore, it is usual to have at least 3 people working on an evaluation; 4 if you include the social statistician. Working in a team, these people discuss with each other. They disagree, explain, argue, and develop ideas. Two important 'Why?' moments arise:

1. Why do you think that? What was the evidence that made you think that?
2. Why are we disagreeing? What is the language / wording / conceptual difference that underpins our disagreement?

When answering these questions, ironing it all out is not the aim. Realising that this is fundamental is important. Retroduction leads to better arguments that explain not only success, but also failure of the intervention in diverse circumstances.

Figure 3 illustrates how retroduction will lead to generating new data, having discussions about our shared premises, and creating other forms of feedback loop.

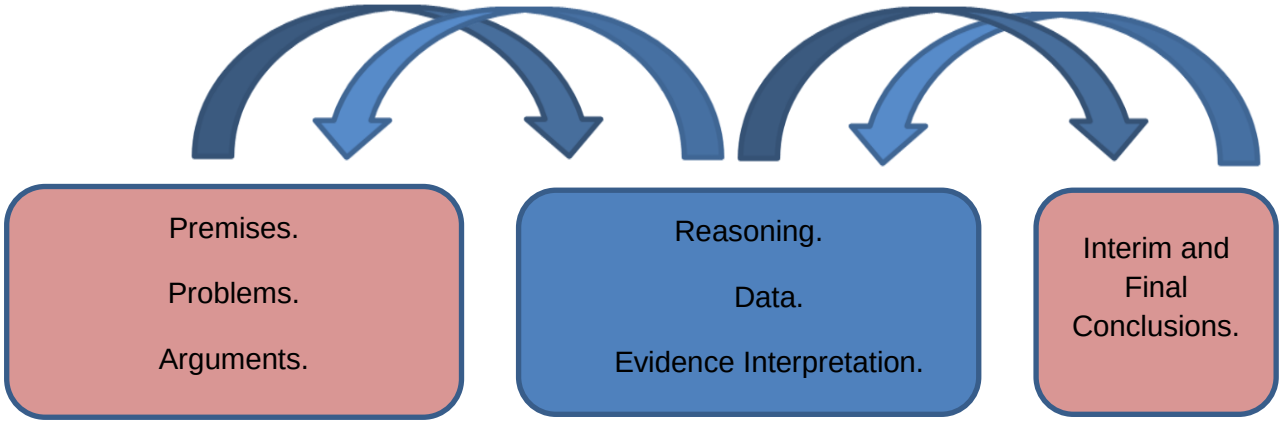

Figure 3: The Evaluation Stages, Showing the Arena Stage and the Feedback Loops of Reasoning & Discussion

Mixed Methods Using Retroduction Is Now Common

The word ‘retroduction’ arose in the 1980s during a debate on ‘scientific realism’. Its historical antecedents arise in a confusing argument among philosophers, which we must avoid (retrodiction being a different thing).

Numerous authors then posited four forms of logic for social research, all arguing that we can combine them (Blaikie, *op cit.* for example). Other authors did not like to admit deductive logic had any purchase, and moved toward a supposed ‘constructivist’ pole. Most authors just try to avoid a sticky argument. But it is really simple. You can do induction for one month, deduction for a few days the next month, then write all that up and do abduction by observing in situ for a few weeks, and then discuss among your team and do retroduction for a week. ‘Doing’ each means creating the input and output stage of each. They are very different. But all are valuable in their own rightful sphere.

Mixed Methods with Multiple Pathways of Cause are Uncommonly Good

What is missing in the RCT statistics is a sense of equifinality. In reality, multiple pathways can lead to the same good outcome. For example, to limit ‘child labour’, achieving high household wealth will work, and achieving a welcoming school with breakfast and noon meals provided free to all children will also work. To take this further, the school may need toilets suitable for both girls and boys as separate blocks. We define this mechanism using a helpful terminology from logic:

- Toilet provision is a necessary condition for one of the sufficient pathways.
- High household wealth is sufficient as a pathway to having no children in ‘Child Labour’.
- Having noon meals at school is not sufficient to eradicate ‘Child Labour’.
- Having both breakfast and noon meals at school, and a welcoming school, combined with the toilet provision, is a sufficient pathway to removing ‘Child Labour’.

Thus, W or (N&B)&T is sufficient for removing ‘Child Labour’ at household level. (Also written $W \cup [(N \cap B) \cap T] \Rightarrow \sim C$, where ~C refers to the absence of child labour .)

Similarly, by blocking the good outcome, obstacles can take many forms. When we think of a ‘bad’ outcome, like ‘child labour’, we need to consider both outcomes:

- Factors affecting having Child Labour
- Factors affecting the eradication of Child Labour.

A large literature shows with case-study methods that multiple pathways commonly occur. Known as ‘equifinality’, this leads to doubts about the closure of the statistical models often used in RCT. If we have some unspecified necessary condition T,

which is part of other apparently sufficient conditions for Y, we may not realise that Y in future will fail, due to the change in T. If we consider the treatment (noon meals and breakfasts) without considering T, yet T is crucial, then we do not have a good arena for the whole conversation about removing Child Labour.

The project should be learning about such factors. A learning community can be created. Ragin has argued for this in a series of books (2000, 2008; and Byrne and Ragin, eds. 2009).

Another example. At high school, to reach high scores, a student may have talent, support at home, a good school, or high resources in their school arising from government funding. Not all of these conditions is required. If such arguments are true – and they often are – then statistical methods that conclude 'no effect' in the face of ambiguity are wrongheaded. Yet it is an empirical question in each situation. The RCT proponents may tend to assume it is a simple empirical question. I would tend to assume it is quite complex. Which factor is sufficient? What condition is INUS? (INUS is a necessary condition as part of a sufficient pathway.) Will we appreciate the difference?

A query to RCT supporters: If something is crucial for some cases, but irrelevant for others, can the data cover it? Yes it can.

A second query to RCT supporters: Do you assume both necessary and sufficient status for every X, A, B, S, and T? Surely not. Statisticians tend to assume if X represents a real cause of Y then it is both necessary and sufficient for Y.

But if X is composite then this is not symmetrically true, even if X is sufficient for Y. In general if X is A&B&S, and X is sufficient for Y, we still have a problem:

With regard to counterfactuals, for such an X, Not-X does not imply (is not sufficient for) Not-Y, even if X is sufficient for Y (see appendix). Nonreversal of causality in the strict case is typical.

In brief, if:

- A= X1&X2&X3 (Eq. 1)
- B = X4&X5 (Eq. 2)
- (A&B) is sufficient for Y (Eq. 3)

Where => refers to 'is sufficient for' and does not mean 'if and only if'.

Then

- ~A => ~(A∩B) =N> ~Y

Where =N> refers to 'is necessary for'.

~A∪~B =N> ~Y (Eq. 4)

The lesson is that if either A or B is absent, it does not guarantee the failure of Y.

Key: ∩ AND
∪ OR
~ NOT
=> is sufficient for
=N> is necessary for
IFF if and only if, which means 'is necessary and sufficient for' and is also based on closure of the model, ie no omitted variables.

Suppose A was Structural background and B was the treatment including an INUS condition,

And we have evidence that together they support the achievement of the outcome,

Then we can't say that the absence of one or the other will cause the failure of the outcome!

This is non-intuitive to most statisticians. They assume that moving upward on a curve is symmetric to moving back downward on that curve. This implies they have conflated => with =N> and IFF. These are three different operators in logic.

Abductive Impact Evaluation is Unlikely to Work Alone But Does Work in Bridging to Action

Mixed methods are very good because the speakers in the abduction and monitoring part tell us what we are getting wrong. Through the practice of good listening, stakeholders and team members will make researchers realise what is important in the mass of data.

The use of anthropological and ethnographic methods is very popular in development evaluation, but the staffing costs have to sit alongside other costs and thus the project managers must be ready to defend and argue the case for 'using' ethnography. At one level ethnography does not sit easily with project evaluation because ethnographers intrinsically want to be open about their findings and not have a closed agenda from the start. Therefore, no promises will be made. However, most development researchers deeply appreciate the multifaceted knowledge that the development community gains from ethnographic practice and the resulting publications. Evidence for this arises from the widespread inclusion of ethnography in the grants already awarded in the growing Global Challenges Research Fund (a UK based initiative of £1.5 billion in research funds linked to development interventions). But does the anthropologist in the team try to bridge to action?

The essence of how this works is through teams of researchers (and team meetings), and through the circulation of knowledge during and after a project through dissemination. A bridge to action can occur even if the individual ethnographer has not planned it. As long as they are involved in communication and dissemination, their work will have an influence. I value this circuit of knowledge. I argue against allowing the abductive investigation to occur without publications and public presentations as follow up. The whole of the academic international community agrees with this position. What is missing in some disciplines is a respect for what abduction offers. By arguing that the learnings from ethnography fit in as claims within warranted arguments, I have shown how teams can invoke abductive logic without resting a whole argument entirely upon that one logic.

Conclusions

The schisms in the methods literature are largely artificial. Instead, the situation is that different groups of authors set up diverse arenas for discussing how events are affected by development interventions. Due to the existence and influence of academic disciplines, the arena offered by some authors is not acceptable to others. This is the case for example when impact assessment is set up too narrowly, with an ontology and theoretical framework from within just one discipline area.

The big challenge is to set up arenas which mix up the disciplines' sensitivities without becoming too wide-ranging for a project to be feasible. For example, one has to allow for basic elements of the wealth or social class; for institutionalised habits or enculturation; for gender and/or other structural elements that underpin inequality; and for aspects of geography in the region where the intervention has occurred. This article argued in favour of multi-disciplinary approaches to impact evaluation across the whole spectrum from medical and social to business and engineering disciplines.

A third issue is the micro nature of some impact evaluations using RCTs. Nested cases always exist, so being too reductionist in data collection might reflect an arena that had little depth. The impact evaluations lack holistic objects such as the government, the social history, organisational types, or the cultural grounding. Many development projects have had successful evaluations, often including both an 'action' part and a survey-based data-analysis part. Each impact evaluation has specific characteristics, and these could limit the usefulness of their findings in other contexts, giving very restricted external validity. Mixed methods could be useful to understand better these specificities across the micro, meso, and macro levels. It is important to recognise the meso level so in a globalising, international world.

Thus a particular form of pluralism is possible in development evaluation. This kind of pluralism has a technical name in the methodology literature: methodological pluralism. I advocate using methodological pluralism across the action research-

systematic research divide (and thus also the qualitative-quantitative divide); and secondly, at the same time, pluralism of disciplines, ie transdisciplinarity.

References

Adams, J., Witten, K., & Conway, K. (2009). Community development as health promotion: evaluating a complex locality-based project in New Zealand. *Community Development Journal, 44*(2), 140-157. doi: 10.1093/cdj/bsm049

Agarwal, B. (2018). Can Group Farms Outperform Individual Family Farms? Empirical Insights from India, *World Development* 108:8, 57-73.

Agol, D., Latawiec, A. E., & Strassburg, B. B. N. (2014). Evaluating impacts of development and conservation projects using sustainability indicators: Opportunities and challenges. *Environmental Impact Assessment Review, 48*, 1-9. doi: 10.1016/j.eiar.2014.04.001

Alfieri, J., Portelance, L., Souhami, L., Steinert, Y., McLeod, P., Gallant, F., & Artho, G. (2012). Development and Impact Evaluation of an E-Learning Radiation Oncology Module. *International Journal of Radiation Oncology Biology Physics, 82*(3), E573-E580. doi: 10.1016/j.ijrobp.2011.07.002

Allmark, Peter, and Katarzyna Machaczek (2018), Discussion Paper: Realism and Pragmatism in a mixed methods study, *J Adv Nurs.*74:1301–1309., DOI 10.1111/jan.13523.

Arnold, B. F., Khush, R. S., Ramaswamy, P., London, A. G., Rajkumar, P., Ramaprabha, P., . . . Colford, J. M. (2010). Causal inference methods to study nonrandomized, preexisting development interventions. *Proceedings of the National Academy of Sciences of the United States of America, 107*(52), 22605-22610. doi: 10.1073/pnas.1008944107

Aus, J. P. (2009). Conjunctural causation in comparative case-oriented research. *Quality & Quantity, 43*(2), 173-183. doi: 10.1007/s11135-007-9104-4

Barrett, C. B., & Carter, M. R. (2010). The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections. *Applied Economic Perspectives and Policy, 32*(4), 515-548. doi: 10.1093/aepp/ppq023

Basinga, P., Gertler, P. J., Binagwaho, A., Soucat, A. L. B., Sturdy, J., & Vermeersch, C. M. J. (2011). Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *Lancet, 377*(9775), 1421-1428. doi: 10.1016/s0140-6736(11)60177-3

Befani, B., Barnett, C., & Stern, E. (2014). Introduction - Rethinking Impact Evaluation for Development. *Ids Bulletin-Institute of Development Studies, 45*(6), 1-5. doi: 10.1111/1759-5436.12108

Blaikie, N.W.H. (2000). *Designing Social Research: The logic of anticipation*. Cambridge, UK ; Malden, MA: Polity Press : Blackwell.

Blaikie, P. (1993). *Approaches to Social Enquiry*. Cambridge: Polity.

Bose, R. (2010). CONSORT Extensions for Development Effectiveness: guidelines for the reporting of randomised control trials of social and economic policy interventions in developing countries. *Journal of Development Effectiveness, 2*(1), 173-186. doi: 10.1080/19439341003624441

Brink, M., et al. (2011). "Sustainable management through improved governance in the game industry." *South African Journal of Wildlife Research* **41**(1): 110-119.

Brunie, A., Fumagalli, L., Martin, T., Field, S., & Rutherford, D. (2014). Can village savings and loan groups be a potential tool in the malnutrition fight? Mixed method findings from Mozambique. *Children and Youth Services Review, 47*, 113-120. doi: 10.1016/j.childyouth.2014.07.010

Byrne, D., and C. Ragin, Eds. 2009. *Handbook of Case-Centred Research Methods,* London: Sage.

Caspari, A. (2009). 'Rigorous' Impact Evaluation - Methodological and Conceptual Approaches for Measuring Impact in Development Cooperation. *Zeitschrift Fur Evaluation, 8*(2), 183-+.

Clemens, M. A., & Demombynes, G. (2011). When does rigorous impact evaluation make a difference? The case of the Millennium Villages. *Journal of Development Effectiveness, 3*(3), 305-339. doi: 10.1080/19439342.2011.587017

Copestake, J. (2014). Credible impact evaluation in complex contexts: Confirmatory and exploratory approaches. *Evaluation, 20*(4), 412-427. doi: 10.1177/1356389014550559

Creswell, J. W. (1994). Research design: Qualitative and quantitative approaches. Thousand Oaks, CA: Sage.

Creswell, John W., and Vicki L. Plano Clark (2018), *Designing and Conducting Mixed Methods Research,* 3rd ed., London: Sage.

de Lange, E., Woodhouse, E., & Milner-Gulland, E. J. (2016). Approaches Used to Evaluate the Social Impacts of Protected Areas. *Conservation Letters, 9*(5), 327-333. doi: 10.1111/conl.12223

Downward, P. and A. Mearman (2007). "Retroduction as Mixed-Methods Triangulation in Economic Research: Reorienting economics into social science." *Camb. J. Econ*. 31(1): 77-99.

Dubowitz, T., Levinson, D., Peterman, J. N., Verma, G., Jacob, S., & Schultink, W. (2007). Intensifying efforts to reduce child malnutrition in India: An evaluation of the Dular program in Jharkhand, India. *Food and Nutrition Bulletin, 28*(3), 266-273. doi: 10.1177/156482650702800302

Duvendack, M., Hombrados, J. G., Palmer-Jones, R., & Waddington, H. (2012). Assessing 'what works' in international development: meta-analysis for sophisticated dummies. *Journal of Development Effectiveness, 4*(3), 456-471. doi: 10.1080/19439342.2012.710642

Fernald, L. C. H., Galasso, E., Qamruddin, J., Ranaivoson, C., Ratsifandrihamanana, L., Stewart, C. P., & Weber, A. M. (2016). A cluster-randomized, controlled trial of nutritional supplementation and promotion of responsive parenting in Madagascar: the MAHAY study design and rationale. *Bmc Public Health, 16*. doi: 10.1186/s12889-016-3097-7

Funnell, Sue, and Patricia J. Rogers (2011). Purposeful Program Theory: effective use of theories of change and logic models. Sue Funnell and Patricia J Rogers, Sydney: Jossey-Bass.

Gelli, A., Becquey, E., Ganaba, R., Headey, D., Hidrobo, M., Huybregts, L., and H. Guedenet (2017). Improving diets and nutrition through an integrated poultry value chain and nutrition intervention (SELEVER) in Burkina Faso: study protocol for a randomized trial. *Trials, 18*. doi: 10.1186/s13063-017-2156-4

Gelli, A., Margolies, A., Santacroce, M., Roschnik, N., Twalibu, A., Katundu, M., . . . Ruel, M. (2018). Using a Community-Based Early Childhood Development Center as a Platform to Promote Production and Consumption Diversity Increases Children's Dietary Intake and Reduces Stunting in Malawi: A

Cluster-Randomized Trial. *Journal of Nutrition, 148*(10), 1587-1597. doi: 10.1093/jn/nxy148
Gimenez, A. and A. Perez-Foguet (2010). "Challenges for Water Governance in Rural Water Supply: Lessons Learned from Tanzania." Water Resources Development **26**(2): 235-248.
Hansen, H., Andersen, O. W., & White, H. (2011). Impact evaluation of infrastructure interventions. *Journal of Development Effectiveness, 3*(1), 1-8. doi: 10.1080/19439342.2011.547659
Hellstrom, E. (2001). Conflict cultures - Qualitative Comparative Analysis of environmental conflicts in forestry. Silva Fennica, 2-109.
Hunt, Sheldon (1994) "A Realist Theory of Empirical Testing: Resolving the Theory-Ladenness / Objectivity Debate", *The Philosophy of Social Sciences*, 24:2.
Ir, P., Korachais, C., Chheng, K., Horemans, D., Van Damme, W., & Meessen, B. (2015). Boosting facility deliveries with results-based financing: a mixed-methods evaluation of the government midwifery incentive scheme in Cambodia. *Bmc Pregnancy and Childbirth, 15*. doi: 10.1186/s12884-015-0589-x
Johnson, N. L., Kovarik, C., Meinzen-Dick, R., Njuki, J., & Quisumbing, A. (2016). Gender, Assets, and Agricultural Development: Lessons from Eight Projects. *World Development, 83*, 295-311. doi: 10.1016/j.worlddev.2016.01.009
Kambala, C., Lohmann, J., Mazalale, J., Brenner, S., Sarker, M., Muula, A. S., & De Allegri, M. (2017). Perceptions of quality across the maternal care continuum in the context of a health financing intervention: Evidence from a mixed methods study in rural Malawi. *Bmc Health Services Research, 17*. doi: 10.1186/s12913-017-2329-6
Kelcey, B., Shen, Z. C., & Spybrook, J. (2016). Intraclass Correlation Coefficients for Designing Cluster-Randomized Trials in Sub-Saharan Africa Education. *Evaluation Review, 40*(6), 500-525. doi: 10.1177/0193841x16660246
Kikuchi, K., Ansah, E., Okawa, S., Shibanuma, A., Gyapong, M., Owusu-Agyei, S., Ghana, E. I. R. P. (2015). Ghana's Ensure Mothers and Babies Regular Access to Care (EMBRACE) program: study protocol for a cluster randomized controlled trial. *Trials, 16*. doi: 10.1186/s13063-014-0539-3
King, E., & Samii, C. (2014). Fast-Track Institution Building in Conflict-Affected Countries? Insights from Recent Field Experiments. *World Development, 64*, 740-754. doi: 10.1016/j.worlddev.2014.06.030
King, Gary, and Richard Nielsen (forthcoming), Why Propensity Scores Should Not Be Used for Matching, *Political Analysis*. Author pre-publication copy at http://j.mp/2ovYGsW .
Lam, W. F., & Ostrom, E. (2010). Analyzing the dynamic complexity of development interventions: lessons from an irrigation experiment in Nepal. *Policy Sciences, 43*(1), 1-25. doi: 10.1007/s11077-009-9082-6
Lan, J., & Yin, R. S. (2017). Research trends: Policy impact evaluation: Future contributions from economics. *Forest Policy and Economics, 83*, 142-145. doi: 10.1016/j.forpol.2017.07.009
Langer, L., Winters, N., & Stewart, R. (2014). Mobile Learning for Development: Ready to Randomise? In M. Kalz, Y. Bayyurt & M. Specht (Eds.), *Mobile as Mainstream-Towards Future Challenges in Mobile Learning, Mlearn 2014* (Vol. 479, pp. 156-167).
Layder, D. (1993). *New Strategies in Social Research*. Cambridge, Polity Press.
Lopez and Scott (2000). *Social Structures*. *

Lubinga, S. J., Jenny, A. M., Larsen-Cooper, E., Crawford, J., Matemba, C., Stergachis, A., & Babigumira, J. B. (2014). Impact of pharmacy worker training and deployment on access to essential medicines and health outcomes in Malawi: protocol for a cluster quasi-experimental evaluation. *Implementation Science, 9*. doi: 10.1186/s13012-014-0156-2

Luo, L. P., & Liu, L. (2014). Reflections on conducting evaluations for rural development interventions in China. *Evaluation and Program Planning, 47*, 1-8. doi: 10.1016/j.evalprogplan.2014.06.004

Masset, E., & Gelli, A. (2013). Improving community development by linking agriculture, nutrition and education: design of a randomised trial of "home-grown" school feeding in Mali. *Trials, 14*. doi: 10.1186/1745-6215-14-55

Maxwell, J., & Mittapalli, K. (2010). Realism as a stance for mixed methods research. In A. Tashakkori, & C. Teddlie (Eds.), Sage handbook of mixed methods in social and behavioural research (2nd ed., pp. 145–167). Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781506335193

McHugh, N., Biosca, O., & Donaldson, C. (2017). From wealth to health: Evaluating microfinance as a complex intervention. *Evaluation, 23*(2), 209-225. doi: 10.1177/1356389017697622

Mock, N. B., Magnani, R. J., Dikassa, L., Rice, J. C., Abdoh, A. A., Bertrand, W. E., & Mercer, D. M. (1993). THE UTILITY OF CASE-CONTROL METHODS FOR HEALTH-POLICY AND PLANNING ANALYSIS - AN ILLUSTRATION FROM KINSHASA, ZAIRE. *Evaluation and Program Planning, 16*(3), 199-205. doi: 10.1016/0149-7189(93)90004-r

Morgan, Jamie, and Wendy Olsen, (2007) “Defining Objectivity In Realist Terms: Objectivity as a Second-Order “Bridging” Concept”, *Journal of Critical Realism*, 6:2, pgs 250-266; republished 2015 by Taylor & Francis, URL https://doi.org/10.1558/jocr.v6i2.250 .

Morgan, Jamie, and Wendy Olsen, (2008) “Defining Objectivity In Realist Terms: Objectivity as a Second-Order “Bridging” Concept, Part 2: Bridging Into Action”, *Journal of Critical Realism,* 7:1, pages 107-132; URL doi: 10.1558/jocr.v7i1.107 See also open access URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.630.7118&rep=rep1&type=pdf

Murshed-e-Jahan, K., Ali, H., Upraity, V., Gurung, S., Dhar, G. C., & Belton, B. (2018). Making sense of the market: Assessing the participatory market chain approach to aquaculture value chain development in Nepal and Bangladesh. *Aquaculture, 493*, 395-405. doi: 10.1016/j.aquaculture.2017.06.003

Nathan, S., Kemp, L., Bunde-Birouste, A., MacKenzie, J., Evers, C., & Shwe, T. A. (2013). "We wouldn't of made friends if we didn't come to Football United": the impacts of a football program on young people's peer, prosocial and cross-cultural relationships. *Bmc Public Health, 13*. doi: 10.1186/1471-2458-13-399.

Ngwenya, Barbara Ntombi Ngwenya, Ketlhatlogile Keta Mosepele and Lapologang Magole (2012), A case for gender equity in governance of the Okavango Delta fisheries in Botswana, *Natural Resources Forum* 36 (2012) 109–122.

Olsen, Wendy (2009b), “Non-Nested and Nested Cases in a Socio-Economic Village Study”, chapter in D. Byrne and C. Ragin, eds. (2009), *Handbook of Case-Centred Research*, London: Sage.

Olsen, Wendy (2012), *Data Collection,* London: Sage.

Olsen, Wendy (forthcoming), “Social Statistics Using Strategic Structuralism and Pluralism”, chapter in Philosophy of Social Science edited volume, *Frontiers*

of Social Science: A Philosophical Reflection, Editor: Michiru Nagatsu and Attilia Ruzzene. London: Bloomsbury Publishing.
Orr, L. L. (2015). 2014 Rossi Award Lecture:* Beyond Internal Validity. *Evaluation Review, 39*(2), 167-178. doi: 10.1177/0193841x15573659
Pollard, S. and D. DuToit (2011). "Towards Adaptive Integrated Water Resources Management in Southern Africa: The Role of Self-organisation and Multi-scale Feedbacks for Learning and Responsiveness in the Letaba and Crocodile Catchments." *Water Resource Management* **25**: 4019-4035.
Pradhan, M., Brinkman, S. A., Beatty, A., Maika, A., Satriawan, E., de Ree, J., & Hasan, A. (2013). Evaluating a community-based early childhood education and development program in Indonesia: study protocol for a pragmatic cluster randomized controlled trial with supplementary matched control group. *Trials, 14*. doi: 10.1186/1745-6215-14-259
Ragin, C. C. (2008), *Redesigning Social Inquiry: Fuzzy Sets and Beyond*, University of Chicago Press.
Ragin, C.C. (2000), *Fuzzy Set Social Science*, *.
Ravallion, M. (2009). Evaluation in the Practice of Development. *World Bank Research Observer, 24*(1), 29-53. doi: 10.1093/wbro/lkp002
Rihoux, B. (2006). Qualitative Comparative Analysis (QCA) and related systematic comparative methods: recent advances and remaining challenges for social science research. *International Sociology*, 21(5), 679-706.
Rihoux, B., and M. Grimm, eds. (2006). *Innovative Comparative Methods For Policy Analysis: Beyond the quantitative-qualitative divide.* New York, NY, Springer.
Sayer, A. (2000) *Realism in Social Science*. London: Sage.
Smithson, M. and J. Verkuilen (2006). *Fuzzy Set Theory: Applications in the social sciences*. Thousand Oaks; London, Sage Publications.
Snow, D. and D. Cress (2000). "The Outcome of Homeless Mobilization: the Influence of Organization, Disruption, Political Mediation, and Framing." *American Journal of Sociology* 105(4): 1063-1104.
Ssengooba, F., McPake, B., & Palmer, N. (2012). Why performance-based contracting failed in Uganda - An "open-box" evaluation of a complex health system intervention. *Social Science & Medicine, 75*(2), 377-383. doi: 10.1016/j.socscimed.2012.02.050
Taft, A. J., Small, R., Humphreys, C., Hegarty, K., Walter, R., Adams, C., & Agius, P. (2012). Enhanced maternal and child health nurse care for women experiencing intimate partner/family violence: protocol for MOVE, a cluster randomised trial of screening and referral in primary health care. *Bmc Public Health, 12*. doi: 10.1186/1471-2458-12-811
Tashakkori, A., & Teddlie, C. (1998). Mixed methodology: Combining qualitative and quantitative approaches (Applied Social Research Methods, No. 46). Thousand Oaks, CA: Sage.
Taylor, A., et al. (2012). "Fostering Environmental Champions: A process to build their capacity to drive change." *Journal of Environmental Management* **98**: 84-97.
Teddlie, Charles, and Abbas Tashakkori (2003) *Handbook of Mixed Methods in Social & Behavioral Research*, Thousand Oaks: Sage.
Teddlie, Charles, and Abbas Tashakkori (2009) *Foundations of Mixed Methods Research*, London: Sage.

The Aspen Institute (2004). Theory of Change as a Tool For Strategic Planning: A report on early experiences. Author Andrea A Anderson. For The Aspen Institute Roundtable on Community Change.

Tremblay, C. and J. Gutberlet (2010). "Empowerment through participation: assessing the voices of leaders from recycling cooperatives in Sa ˜o Paulo, Brazil." *Community Development Journal* **47**(2): 282-302.

UK Aid Connect (2018), *Guidance Note: Developing a Theory of Change*. Downloaded January 2019, URL https://assets.publishing.service.gov.uk/media/5964b5dd40f0b60a4000015b/UK-Aid-Connect-Theory-of-Change-Guidance.pdf.

UNDP/Hivos (2011). *Theory of Change. A Thinking and Action Approach to Navigate in the complexity of social change processes.* Author Iñigo R Eguren. For Hivos, The Netherlands, and the UNDP Regional Centre for Latin America and the Caribbean.

White, H. (2013). An introduction to the use of randomised control trials to evaluate development interventions. *Journal of Development Effectiveness, 5*(1), 30-49. doi: 10.1080/19439342.2013.764652

Williamson, M., Cardona-Morrell, M., Elliott, J. D., Reeve, J. F., Stocks, N. P., Emery, J., . . . Gunn, J. M. (2012). Prescribing Data in General Practice Demonstration (PDGPD) project - a cluster randomised controlled trial of a quality improvement intervention to achieve better prescribing for chronic heart failure and hypertension. *Bmc Health Services Research, 12*. doi: 10.1186/1472-6963-12-273

Appendix.

In Boolean logic, inclusion refers to the subset relation of the observed values of X and the observed values of Y, where X and Y are set membership scores. It is common to think of X<Y implying that X is a subset of Y. The values being set membership scores, X being less than Y for all cases $\{x_i, y_i]$ is what we mean by X being included in Y.

It might be obvious that the **inclusion** of X in Y is non-reversable with the inclusion of Y in X. Ragin has argued that if X is a subset of Y, also known as inclusion as defined by Smithson and Verkuilen (2006), then X is sufficient for Y and Y is necessary (*sic*) for X. In notation:

If X => Y then Y =N> X (Eq. 0, a statement of the transitivity of the subset relation)

Ragin would say if X is a subset of Y then Y is a superset of X. For Smithson and Verkuilen this would be expressed in terms of inclusion and non-inclusion.

Another rule we have in logic is that if X => Y then ~Y => ~X (Eq. 0.1)

Applying both rules we reach a third conclusion: if X => Y then ~X =N> ~Y. (Eq. 0.2)

(Specifically I have applied the rule in Eq. 0.1 to the last expression in Eq. 0.2.)

Importantly, in such a case as Eq. 0.1, the absence of X is not sufficient to cause the absence of Y. In fuzzy sets, the absence of X would be either X=0 (crisp measurement) or X≤0.5 (fuzzy measurement).

Based on background knowledge of reality, we will know which statement to argue for. In many research projects Ragin has placed all the information available over a time period into one measurement set, thus consolidating change-over-time to the overall existence of situations; and he then looks instead at overall results using hindsight. Based on such evidence, the differences between 'if and only if' (IFF), sufficiency, and necessity are crucial. Ragin does not avoid causal language, whereas Smithson and Verkuilen avoid it entirely (2006). Ragin has established that arguing for necessity and sufficiency of cause at the same time can be unwise. Unpicking them can lead to new knowledge about equifinality and the role of contextual factors.

Nonreversal of causality in the strict case is typical. We know from logic that if X is sufficient for Y, then Y is necessary for X. We cannot conclude from X => Y that ~X is sufficient for ~Y. Instead the following logic is best followed to reach a sound conclusion for multiple causes.

If we know that two sets of factors are both, when combined together, sufficient for an outcome Y, we can write:

- A= X1&X2&X3 (Eq. 1)

- B = X4&X5 (Eq. 2)
- (A&B) is sufficient for Y (Eq. 3)

Note the key, where we use specific notation for sufficiency; => refers to 'is sufficient for' and does not mean 'if and only if'.

Then

- ~A => ~(A∩B) by the definition of 'Not'.
- ~(A∩B) =N> ~Y by the rule expressed in Eq. 0.2.

Key:
∩ AND
∪ OR
~ NOT
=> is sufficient for
=N> is necessary for
IFF if and only if, which also means 'is necessary and sufficient for'

Noticing in particular that =N> refers to 'is necessary for', and thus that it implies only that ~(A∩B) is something that universally accompanies ~Y, but it does not assure that ~Y occurs. (For example, to eradicate child labour, ie achieve ~Y, you may have to do more than get rid of (A∩B).) The key conclusion is that either A or B may be missing, yet in itself that does not inform us much about what will happen to Y!

~A∪~B =N> ~Y (Eq. 4) by the application of deMorgan's law to the lefthand side.

DeMorgan's law is explained in Ragin (2000).

The lesson is that if either A or B is absent, it does not guarantee the failure of Y.

Suppose A was structural background and B was the treatment including an INUS condition,

And we have evidence that together they support the achievement of the outcome,

Then we can't say that the absence of one or the other will cause the failure of the outcome!

My conclusions have depended upon a key assumption that the world is an open system and not a closed system. The IFF relation assumes closure, ie that the data cover all relevant variables. By contrast, the relations => and =N> (sufficiency and necessity) work without making this assumption. Ragin's many books thus do not assume away complexity, whereas the RCT literature and the mathematicians Smithson and Verkuilen (2006) tend to argue that the variables exhaust the relevant evidence. Interesting reflections on these issues are found in Barrett and Carter (2010) and Duvendack, et al. (2012).

[i] Two cases can be considered 1) if an attempt is made at fully deductive mixed-methods research, with a single 'logic', this attempt will contradict itself. As explained by Hunt (*), the theoretical framework 'chosen' will imply other logics and other practical engagements in the world, not acknowledged in the 'hypothesis, test, conclude' deductive logic. 2) The second case is that the mixed-methods research would be multi-logic and include a randomised controlled trial. Here the part which is RCT would use deduction, while other parts

would use retroduction and induction, carry out synthesis, and respond to feedback. However, if this pollutes the RCT, then the methods chosen as a mix contradict the clarity of the RCT control group separation from the treated groups. This is why I was concerned about the RCT method within mixed-methods development research.