May 2022

CEDIL Methods Working Paper 8

The unfinished evidence revolution: riding the five waves

Howard White

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL methods working papers

The CEDIL methods working paper series offers innovative research methods to develop impact evaluation and evidence synthesis work in low- and middle-income countries.

About this working paper

This paper by Howard White, titled ‘The unfinished evidence revolution: riding the five waves’ examines the current state of each of the waves of the evidence revolution, reflecting on how the evidence revolution has developed through different waves, examining the current state of each wave, before discussing the implications for evaluation in the field of international development.

The content of this paper is the sole responsibility of the authors and does not represent the opinions of CEDIL or the Foreign, Commonwealth & Development Office. Any errors and omissions are also the sole responsibility of the authors. Please direct any comments or queries to the corresponding author, Howard White, h.white@campbellcollaboration.org

Suggested citation: White, H. (2022) *The unfinished evidence revolution: riding the five waves*. CEDIL Methods Working Paper 8. London/Oxford: Centre of Excellence for Development Impact and Learning (CEDIL). https://doi.org/10.51744/CMWP8

Cover design: PhilDoesDesign (cover needs updating for every working paper)

CEDIL methods working paper: The unfinished evidence revolution: riding the five waves

Authors

Howard White, Campbell Collaboration

Contents

List of tables, figures, and boxes

Introduction

The last thirty years have seen an evidence revolution which has developed in four waves: outcome monitoring, impact evaluation, systematic reviews and knowledge brokering (see White, 2019). But this is an unfinished revolution. Further work, and investment, is needed at all levels of the knowledge pyramid, which are partly represented by different waves of the evidence revolution.

This paper examines the current state of each of the waves of the evidence revolution. I offer a critical perspective on each wave, and discuss the implications for evaluation in the field of international development. Whilst the use of evidence has grown, underuse and misuse remain. More importantly, there has been a failure to institutionalize the use of evidence beyond the intervention level. The last part offers some concluding comments.

What to do with the continuing first wave? Abuses and uses of monitoring data

Outcome monitoring and results frameworks: abuses of monitoring data

The first wave of the evidence revolution was ushered in by the rise of New Public Management (Manning and White, 2014). New Public Management sought to run the public sector more like the private sector, which included a focus on being 'results-oriented'. Results were identified with high-level welfare measures also called outcome or impact measures.[1] The focus on final outcomes was a positive step, replacing an earlier focus of much monitoring solely on inputs, activities and possibly outputs. But the downside was an increased focus on the accountability function of monitoring at the expense of lesson learning.

The adoption of various forms of outcome monitoring became widespread and continues to be common. These days being outcome monitoring is mostly incorporated into 'results frameworks'. Results in this terminology are just the data on the outcomes of interest. But this is simply monitoring outcomes before versus after with no impact evaluation design to determine the extent to which the observed changes can be attributed to the intervention. However, the very word 'results' - and especially speaking of an agency's results - give the impression that there can be some attribution.

Of course, there are many examples of when the change in outcomes in project areas suggested an intervention was more successful than it actually was. A prominent example is the Millennium Villages Project, in which an early analysis of Clemens and Demombynes (2011) showed the project to be having rather less impact than claimed when trends in outcome indicators were compared with those in non-project areas. One of two subsequent

[1] Monitoring and evaluation distinguishes outcomes as being more immediate effects of an intervention and impacts as longer run ones. However, since impact has a different meaning when used in the context of counterfactual impact evaluations the use of the term impact can lead to unnecessary misunderstandings (White, 2009).

impact evaluations of MVP finds only small or no effects on main the main welfare indicators (Masset et al., 2020), and a second finds positive effects on health and agriculture, but not on poverty and nutrition (Mitchell et al., 2018). Similarly, great claims were made for the Bangladesh Integrated Nutrition Project based on substantial reductions in child malnutrition in project areas. But a rigorous evaluation found little or no effect, as such trends were present over the whole country (White and Masset, 2007).

Hence results frameworks may be said to be misleadingly named as it supports a sleight of mind to imply that they enable measurement of results of what an intervention or agency has achieved. 'Results frameworks' lull managers into thinking they have a system of feedback on programme impact whereas in fact they have no such thing.

Uses of monitoring data

This is not to say monitoring is unimportant. Monitoring matters. But the focus on 'results' has given monitoring an accountability role at the expense of lesson learning. Despite the fact that M&E became MLE about a decade ago, the L is largely silent. Monitoring offices complete 'results framework' data used for accountability not lesson learning. Outcome-level monitoring data – and outcome data more generally – help identify problems such as prevalence and variations in prevalence across sub populations – but not what to do about those problems. There are often few if any feedback loops into project management.

Hence, monitoring needs to return to its traditional role which is mainly focused on indicators lower down the causal chain. It would be helpful to have a greater emphasis on the role of monitoring data in project management. This is the traditional use of monitoring data. Ensuring that project managers have information on whether things are happening as planned – equipment delivered, numbers of people recruited and trained, number and volume of microcredit loans made and so on – is important and should not be lost sight of.

This is not to say that opportunities should not be sought to use routine monitoring data for impact evaluations. This is possible for an A/B design in which there are two treatment arms. A rapid impact evaluation can be an adaptive learning approach. Different approaches could be used to promote adoption of the promoted technology to see which achieves higher adoption rates. For example, indoor residual spraying in Mali to treat mosquitos is typically promoted by door-to-door mobilization in villages. A new project considered text messaging as an alternative, possibly cheaper, means of mobilization. However, a rapid RCT with an A/B design showed that the mobile-phone based approach was both less effective and more costly than door-to-door visits. Hence the existing approach was retained (Mangam et al., 2016).

It may also be possible to assess impact of an intervention using monitoring data in the absence of a control group. This approach, common in many settings such as managing factory production processes, is statistical process control which may be used to measure the response to a change in project practice. The impact of such changes may be formally tested using interrupted time series analysis (ITS), which is a regression-based approach to determine whether the change in practice coincided with an improvement in outcomes

(Fretheim and Tomic, 2015). For example, ITS was applied to evaluate the impact of adopting pre-operative antibiotic prophylaxis at a single hospital in Kenya, finding it led to a reduction in infections of wounds, and the total time and cost of administering antibiotics (Aiken et al., 2013).

Strengthening the base of the knowledge pyramid

This misuse of monitoring data has a more recent manifestation in the conflation of data and evidence. People say such things as 'we can look at the data to see what works'. Data are not evidence. Data are the base or bedrock the 'knowledge pyramid' (see Figure 1). The pyramid shows the successive stages through which evidence is processed or translated to become usable knowledge. The pyramid is not a hierarchy, as is the traditional evidence pyramid with systematic reviews and RCTs at the top. Rather, in the knowledge translation pyramid, evidence is more heavily brokered as we move up the pyramid.

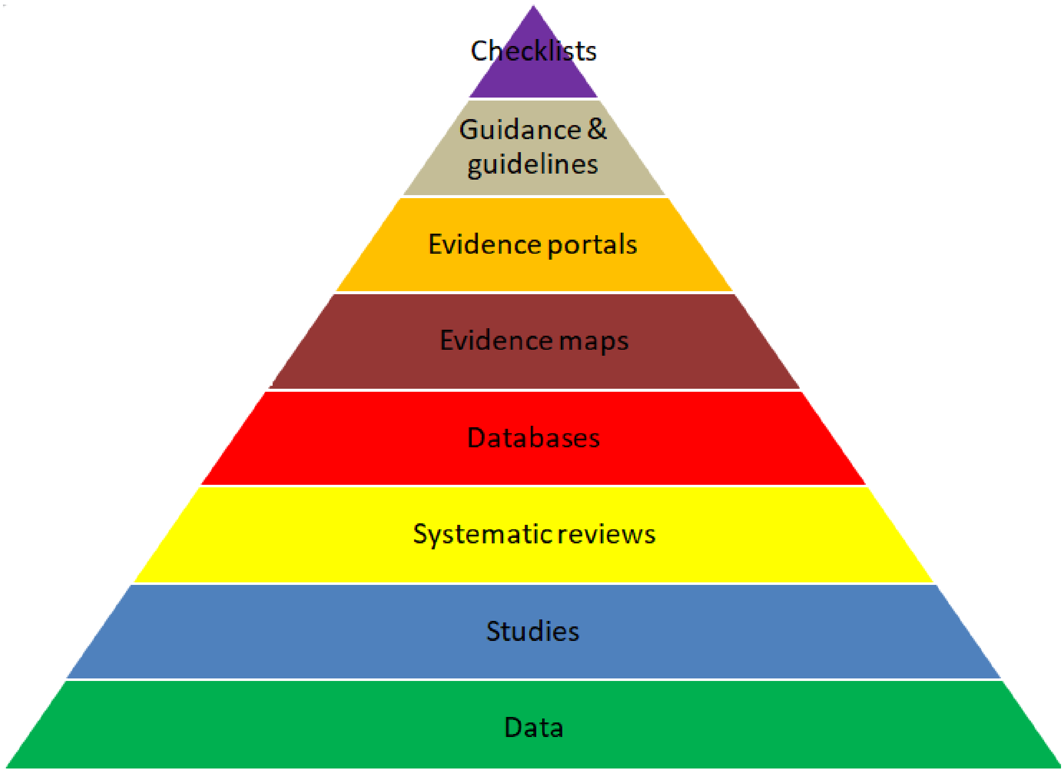
First data are summarized and analyzed with the findings presented in studies and reports. These studies are summarized in reviews. Both studies and reviews may be shown in maps. All these resources can be accessed through evidence platforms. The top three layers of the pyramid are evidence products to inform decisions by decision-makers without the decision-makers having to consult the underlying research. All of these approaches are discussed further below.

But even when the proper role of data is recognized, there are some issues to be addressed in the use of monitoring data.

The first two issues are the related issues of data quality and indicator definition, both of which can be helped by working towards a common set of indicators as is being done in health through the COMET Initiative.[2] Considerable investment and experience has gone into developing modules and questions of the sort used in Living Standards Measurement Surveys (LSMS) and Demographic and Health Surveys (DHS). Researchers need not reinvent the wheel with their own questions. They are often too ready to do so, even though they may be ignorant of the whole research field of questionnaire design (construct validity, validation in different contexts etc.). It is far better for researchers to use existing, validated measures where possible, both for reasons of data quality and comparability between studies. This is a well-developed field in some disciplinary areas, notably psychometrics (psychological measurement), but not properly recognized as an area of research in itself across most development research areas.

[2] http://www.comet-initiative.org/.

Figure 1 Evidence pyramid



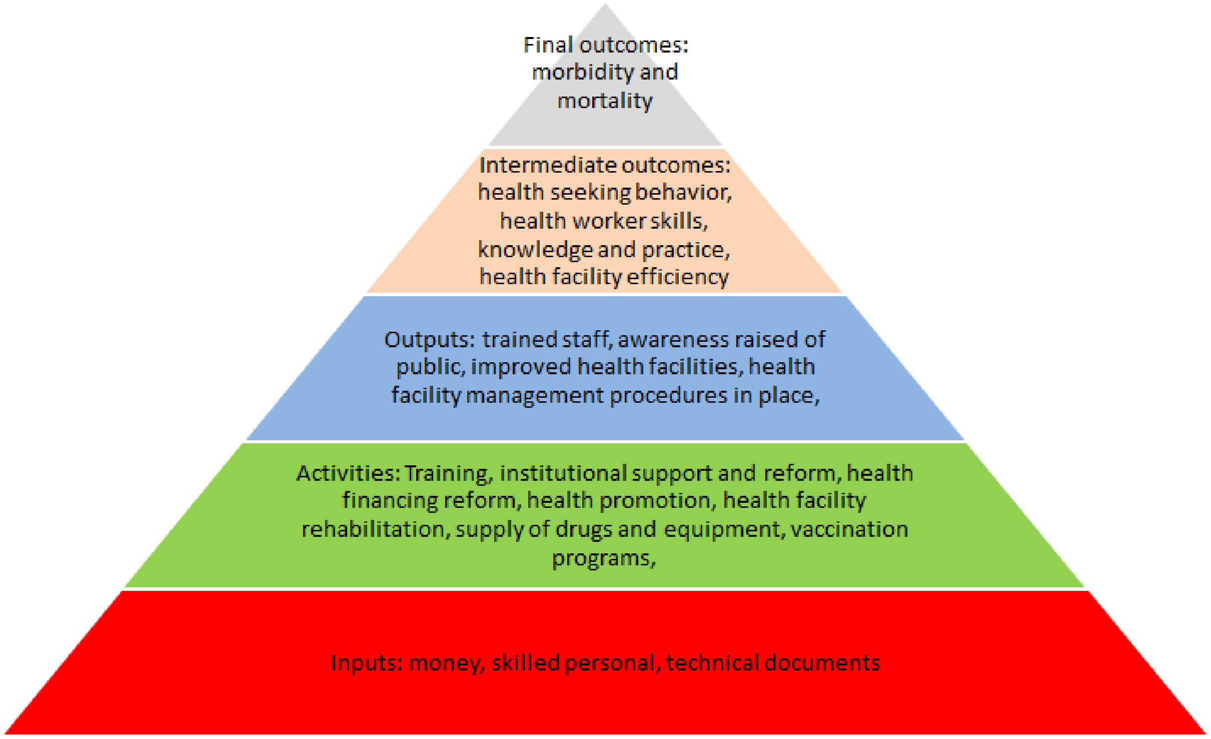
Source: author.

In international development the Millennium Development Goals (MDGs) and the successor Sustainable Development Goals (SDGs) offer a set of international performance indicators and have given a boost to measurement of common indicators. The SDGs have 17 goals, 169 targets and 230 indicators which were selected from over 1,000 proposed indicators. For most (but not all) of these indicators there are agreed methodologies for their measurement, though systematic validation of these measures is lacking. Many of these indicators can be adopted or adapted for intervention level M&E. Agencies which develop results frameworks should align them with the SDG indicators. Project-level results frameworks should use these same indicators where appropriate, which can be done with reference to the 'indicator pyramid' (Figure 2).

The indicator pyramid, is a useful tool in achieving standardized indicators. It recognizes that intervention-specific indicators may be needed lower down the causal chain. But at the upper end of final outcomes and long run impacts common indicators can be used across interventions. The example in Figure 2 demonstrates the concept with health interventions: activities and outputs are very diversified, whereas intermediate and final outcomes are far narrower, applying across a range of interventions. International development organizations, working in partnership with relevant global partnerships, and drawing on the SDGs, should develop indicative sets of indicators by intervention class.[3]

[3] A separate issue is that there is no standard set of interventions in different sectors such as WHO'S International Classification of Health Interventions (ICHI): https://www.who.int/classifications/ichi/en/. This is another area of work for international development organizations and global partnerships.

Figure 2 The indicator pyramid


Final outcomes: morbidity and mortality
Intermediate outcomes: health seeking behavior, health worker skills, knowledge and practice, health facility efficiency
Outputs: trained staff, awareness raised of public, improved health facilities, health facility management procedures in place,
Activities: Training, institutional support and reform, health financing reform, health promotion, health facility rehabilitation, supply of drugs and equipment, vaccination programs,
Inputs: money, skilled personal, technical documents

Source: author.

Standardizing indicators, and recognizing the appropriate use of each, will address two other issues: indicator proliferation and the missing middle.

Excessive reporting requirements, including indicator proliferation. was at the heart of the critique of results by Andrew Natsios, who reproduces an amusing, perhaps apocryphal letter from Wellington to the Foreign Office during his campaign in the Iberian Peninsula (Natsios, 2010). Wellington pointed out that he could keep track of the petty cash and marmalade ration to satisfy Whitehall's form filling demands, or defeat the French, but he could not do both. Could they please decide, which it was to be? But as his enemy, Napoleon, remarked an army marches on its stomach. The marmalade mattered. However, if Wellington was himself overseeing counting the jars each morning he was a terrible micromanager who needed to learn how to delegate. Different people in an organization need different information. The World Bank President needs to know the World Bank's annual spending, the percent of projects performing satisfactorily, what response is prepared for the next health or financial crisis and so on, but an individual Task Manager needs to know the disbursement lag for the specific project he or she is managing. Someone should have been counting the marmalade, but Wellington's own Management Information System needed just two pieces of information – where are the French and how many are they?

Clearly, Wellington would be keen to know the strength and location of his opponents. These indicators thus met a basic criterion for indicators which is to be useful for the intended user.

Indicator proliferation can arise also because people forget this, and start to generate indicators for indicators' sake. We should not collect data which will not be used. So a good monitoring system benefits from user engagement in identifying indicators, ensuring that users can easily access the data in a timely manner and that they have means of acting in ways to influence these indicators.

Identifying which indicators are most relevant is assisted by a well-developed theory of change or logic model. The use of logic models took hold during the 1980s and nineties. During the 2000s practice has shifted to theories of change which are seen as a more useful approach. The extent to which logic models and theory-based approaches differ is a subject to discussion.

One possible difference is the focus on underlying assumptions. Interrogation of the theory of change identifies critical assumptions, that is things which need to be true for the intervention to work as intended, and so which indicators managers should keep an eye on to ensure they are indeed true. So, rather than simply the number of farmers trained in an agricultural extension programme, it is also useful to know whether farmers acquire and use the knowledge being promoted. So, the theory of change informs indicator selection and data collection (Kusek and Rist, 2004). However, a defender of logic models can point out that the logic model approach includes 'risk factors' which are the same as the assumptions of what needs to be in place for the intervention to work.

A second possible difference is the focus on causal processes. The rise of mid-level theory is increasing the focus on explicitly identifying and spelling out the causal process, rather than leaving it as an unnamed link from one box to another in the theory of change. This discussion has also focused attention on the different types of causal link (see Davies 2018, and White 2021). The lack of attention to causal processes in many applications of logic models might mean that in fact there is at best a weak link from inputs and activities to outputs and outcomes.

DFID undertook a major realignment of its aid strategy to be more poverty-oriented in the late 1990s. Shortly after that happened, David Booth and I were asked to review the new generation of country strategy papers. Here we identified what we called the problem of the missing middle,[4] which soon became more widely recognized in the results community. That is, DFID project and country level documents had introduced high level poverty indicators, but there were no intermediate indicators supported by a theory of change for how the planned interventions would help achieve those outcomes. In some cases, any such theory would have been very tenuous, as for example support to public sector retrenchment for which the poverty objective was reducing maternal mortality.[5] Since that time, use of the theory of

[4] The findings of the report are incorporated in the critical review of the MDGs in Black and White (2002). The original internal paper for DFID was widely drawn on in the sixth report of the parliamentary International Development Committee, which called its concluding section 'the missing middle' (House of Commons, 2002).

[5] Maternal mortality is anyhow a very bad indicator to use for monitoring as it requires very large sample sizes and is usually only reported with a large time lag. And even then, it is of questionable

change approach has become more widespread, which should lead to wider use of intermediate indicators. Yet one can still come across both monitoring systems and evaluation designs which lack these, partly because of a continued misplaced belief that outcome monitoring tells us about impact.

Failure to adopt a theory of change can result in a missing middle resulting in both a lack of information about intermediate indicators which are useful to signal problems in a timely manner, and also collection of data on final outcomes which are at best loosely related to the intervention.

Implications for evaluation

Monitoring needs to be rescued from the results agenda. Whilst outcome data are being wrongly heralded as measures of impact, attention is being distracted from the proper use of monitoring feedback to manage project performance. This problem is exacerbated by responsible staff lacking skills in designing monitoring systems, using inappropriate tools or indicators, lack of timely reporting, and management simply not having access to, or interest in, monitoring data. All these points, the last in particular, suggest that incentives do not reward project performance.

There are several things the evaluation community, notably evaluation departments, can do to flag the deficiencies in the uses and abuses of monitoring data and contribute to a discussion as to how to address these deficiencies.

These activities include: (1) evaluations of the design and use of the agency's results system including the appropriateness of the indicators and their alignment with the international system (and support to that international system); (2) regular reviews of the design and use of monitoring in projects supported by the agency; (3) consider adopting the World Bank's IEG approach of reviewing self-evaluations including design and use of the project M&E system.

The second wave: the rise of impact evaluation, especially RCTs

In the mid-90s Soniya Carvalho and conducted an assessment of 67 World Bank projects which had been designated as targeting the poor, having been tagged as part of the Programme of Targeted Interventions (PTIs) (Carvalho and White, 1996). Our assessment of the impact of these projects had to rely on performance ratings. We tried to aggregate across indicators for projects in similar sectors, but most projects reported indicators sporadically if at all, with no consistency in indicators across projects. Our report discussed the requirements for rigorous impact evaluation, pointing to promising developments such as the then recent paper by John Newman, Paul Gertler and Laura Rawlings reporting a number of randomized controlled trials (Newman et al. 1994). We ended our discussion will a call for task managers to pay more attention to evaluation design requirements in project design,

accuracy; see, for example, the range of estimates in the review of maternal mortality trends by Alkema et al. (2016).

notably baseline data including a valid comparison group. This recommendation was timely but premature as it was a few years before impact evaluations started to take off.

The rise of impact evaluation was a response to the growing recognition of the bankruptcy of outcome monitoring as a means of measuring program effectiveness. Several factors fueled this movement. On the academic front both JPAL and IPA have been active proponents. Within the World Bank the Chief Economist took the initiative to create the Development Impact Evaluation DIME. And the think tank the Centre for Global Development (CGD) published *When Will We Ever Learn?* (Levine and Savedoff, 2006) shaming bilaterals to join foundations in supporting the creation of the International Initiative for Impact Evaluation (3ie).

In consequence of these initiatives, over 5,000 impact evaluations of development interventions have been published in the last 20 years.[6] These studies are a rich source of information about what works. There have been a range of criticisms such as ethics and cost. I believe these criticisms are largely invalid, and have dealt with them elsewhere (White, 2013). But there are issues in both production and use of RCTs which I address here.

On the production side there has been a failure to always answer the evaluation questions of interest. This is partly because many impact evaluations focus solely on impact, eschewing questions of relevance, implementation, cost and sustainability. Piciotto (2014) has suggested that impact evaluations are not evaluations at all as they do not address these questions. But even on their own terms, many impact evaluations fail to provide information of great use to decision-makers.

This situation is started to change somewhat with attention to second generation questions. First generation questions address whether an intervention works or not. But decision-makers often want to know how to do something, not what to do. Second generation questions explore issues to inform design. For example, an evaluation of a conditional cash transfer may explore the effect of size and timing of transfers, payment mechanism, and of types of condition.

Problems also arise because of a methods bias, with a focus on methods rather than content. This bias surfaces in the context of researcher capture; that is that researchers are interested in producing papers publishable in peer reviewed journals not ones which answer questions of practical relevance. The two are not always the same. But there are also three important design issues. First, there has been a focus on first generation what works questions rather than second generation design and implementation questions. Second, purely quantitative large n designs cannot address all causal questions of interest. Integrated mixed method designs are needed to address issues around implementation, barriers and facilitators of participation, and so on. The use of such designs is increasing but remains too rare, partly as

[6] Source: 3ie impact evaluation repository https://www.3ieimpact.org/evidence-hub/impact-evaluation-repository.

journals in some disciplines, notably economics, will not publish such research.[7] Finally, impact evaluations are best suited to evaluating a single intervention or a small number of combinations of such interventions. They cannot be used to assess a whole project of the sort supported by many international development organizations. Nor can they assess the effectiveness of some classes of intervention such as institutional development. Many of these fall under the heading of complex interventions. There remains a challenge to agree on credible approaches to assessing causal relationships for complex interventions especially for interventions for which small n study designs are necessary. I briefly discuss the last of these.

There are two issues with respect to small n designs: explicit attention to methods, and avoiding biases. Case studies are central to small n evaluation designs. There are a well-established methods for addressing causality in case studies (e.g. Byrne and Ragin, 2013). Blatter and Haverland (2014) identify three approaches: co-variational analysis, process tracing and congruence analysis. These approaches are mostly based on inductive logic rather than the deductive logic of RCTs.[8] This does not offer the same level of causal evidence – in Nancy Cartwright's words, they 'vouch for' the causal link rather than 'clinch' it (Cartwright 2011). That limitation needs to be recognized, but they are the methods we have. There is also a growing methods literature about applying small and middle n methods, notably Ragin (2000), and Goetz and Mahoney (2012), and specifically on process tracing by Beach and Pedersen (2013) and Bennett and Checkel (2014). However, it remains the case that the majority of small n-evaluative reports do not pay explicit attention to methods. Where there is a discussion of methods it more commonly describes methods of data collection not data analysis.[9]

Explicit ex ante attention to methods may help reduce bias, but that alone is unlikely to be sufficient. There are many biases which are likely to influence qualitative evaluations of causal effects towards reporting positive findings (White and Phillips, 2012).

One prevalent source of bias is to adopt an effects of causes approach to the evaluation rather than causes of effects. A effects of causes approach is to pose evaluation questions as 'we did this, what happened as a result?'. A causes of effects approach poses questions as 'this happened, what explains it'. At its crudest, a key informant interview for the former approach starts as follows 'I am working for Agency X to conduct an evaluation of Project Y. What has been the impact of Project Y?'. Such an approach suffers from courtesy bias, as well as funding renewal bias and self-importance bias if speaking to a representative of a local implementing agency. The most important bias is probably simply ignoring other possible explanations of changes which may otherwise be attributed to the project. Amongst small n

[7] Some time ago I was referee for an excellent paper submitted to a leading development journal which had a rich description of the project and its context. The editor's decision went with a minority referee comment to remove all the contextual information and focus on testing the model.

[8] Process tracing involves both deductive and inductive reasoning (see Beach and Pedersen, 2013).

[9] I recently reviewed of ILO evaluations. Of over 100 evaluations assessed the methods section from only seven described an evaluation method rather than data collection.

approaches, general elimination methodology has an explicit focus on considering other possible explanations.

There are at least two qualitative approaches which adopt a causes of effects approach. The first is outcome harvesting, which developed out of outcome mapping. Outcome harvesting engages key stakeholders to identify what they think have been their main achievements, and then constructs and tests a causal chain from the intervention to the outcomes. A possible bias here is that adverse effects may not be identified in this approach. The Qualitative Impact Protocol (QuIP) takes further steps to avoid bias. Interviews are explicitly oriented around 'this change happened, what may have caused it?' (see Copestake et al., 2019). Furthermore, neither the interviewer nor the respondent are aware of the intervention being evaluated.[10] If it was an important factor in the changes it is meant to have affected then it will come up.

In most disciplines using qualitative analysis to answer causal questions – such as history – the causes of effects approach is common; e.g. what were the causes of the Russian Revolution? Similarly, evaluations they should document the policy or institutional change which has taken place and so analyze the likely causes of those changes, and then identify possible channels by which the intervention may have had effects.

There have been limitations in the use of impact evaluations, only part of which stems from their design limitations which have been briefly described above. There are also examples of underuse and misuse.

Underuse: most agencies have no formal institutional linkage between the findings from impact evaluations and their agency's performance measurement system (White and Manning, 2014), or other ways in which those findings may be used. For example, a review by IEG found that half the impact evaluations supported by DIME had not been cited in World Bank documents (IEG, 2012: 94). There are several reasons for this, such as the applicability of RCTs and (potential) coverage, the question that RCTs respond to (what, not how or why), that impact evaluations only covering minority part of complex project, and that the impact evaluations are often done by researchers, disconnected from the institutional decision-making machinery

Misuse: RCTs have high internal validity but are often criticized for having weak external validity. I think these concerns are over-stated, and there is work in progress on many fronts to address the issue. In brief, we know that impact can vary by population and context, and will be affected by intervention design and implementation. The issue is to identify which of these factors matter for impact, and whether these necessary conditions are present in another setting so that the study findings transfer to that setting. Such analysis is best conducted in systematic reviews. Thus, evaluation findings from a study of a specific intervention are appropriately used to inform decisions about that intervention. But to inform

[10] This approach is labelled blindfolding rather than blinding, since the blindfold can be removed at any point. In key informant interviews I have conducted, in which I also started with the change being investigated, I would start to prompt about the role of the agency being evaluated in the last part of the interview if it had not been mentioned at all.

strategy decisions, or programme design decisions elsewhere, it is more appropriate to consult the body of evidence.

However, within the World Bank, and most other development agencies, there is no process for accumulating bodies of evidence from studies to inform strategy, policy, and project design and implementation. The failure to accumulate evidence of what works in a systematic way is a failing across the global funds, and of the MDBs whose rationale is to support global public goods such as the evidence architecture.

Implications for evaluation

The rise of impact evaluation may seem irrelevant for most evaluation departments which conduct *ex post* evaluations, mainly at country, sector and thematic levels. But, in fact, two important issues are raised: the unit of analysis, and the confidence which can be placed in causal statements in the type of evaluations typically conducted by institutional evaluation departments.

It is often pointed out that impact evaluations will typically cover just one component of a project, not the whole project. This is seen as a shortcoming of the approach. The shortcoming is partly that we evaluate what is easy of evaluate which may leave important questions unaddressed. But this issue is not unique to RCTs. In many cases although it is stated that the project is the unit of analysis for evaluation but that is not really so.

In sector and thematic evaluations a universe of projects is defined, and the evaluation summarizes the experience of those projects. But anyone who has conducted such an evaluation is aware of the uncertainties and compromises which have to be made. Sometimes the different components of a development project fit together in one synergistic whole (all fitting in a common theory of change), but more often they are quite discrete activities with different outcomes. And in either case there are variations in performance between components – some bits may work but not others.

Any evaluation which addresses a country, sector or theme necessarily embraces a range of varied projects. Even if an evaluation is concerned with a specific intervention, say conditional cash transfers there will still be considerable heterogeneity of intervention designs and types of support. It is not clear what meaningful evaluation findings for some questions, such as relevance and effectiveness, can be given for a portfolio of projects encompassing many different interventions. There is a lot to be said for restructuring the approach to sector and thematic evaluations to be more focused on evaluating specific interventions. Currently many qualitative evaluations do a very poor job of describing the interventions being evaluated (some do not do it at all!), and the so-called theories of change are not used in any meaningful way to frame the evaluation.

The growth of RCTs has focused attention on attribution, i.e. causal statements. But the fact is that, despite ‘impact’ being part of the DAC criteria routinely used by evaluation agencies, there has been little explicit attention in the evaluation community at large to establishing causality when large n study designs cannot be used. The issue is discussed – for example in the NONIE Guidance (Leeuv and Vaessen, 2009) and the Stern et al (2012) paper commissioned by DFID. More recently, the DFID-funded CEDIL project and a work stream of the official German development evaluation agency, DEval, are concerned with causal evaluation for complex interventions.

These efforts are not working in a vacuum. As discussed above, there is a relevant literature on these issues, which is taking time to get drawn upon in the evaluation community. In our review of small n impact evaluations (White and Phillips, 2012), we commented on the fact that many papers simply say that they will use 'mixed methods' to establish causality, which is insufficient in our view. As argued above, there should be explicit attention to methods and reducing biases, including by adopting a causes of effects approach when using qualitative methods to address causal questions.

This approach is likely to mean evaluating classes of intervention rather than projects which comprise many different interventions. Such studies would be a much better basis for the lesson learning function of accumulating bodies of evidence.

Third wave: Rise of systematic reviews

As the amount of evidence available grows it makes sense to summarize that evidence. This has traditionally been done by literature reviews. But the sources of bias in such reviews are increasingly documented, such as being selective in both which studies are included and the findings from those studies.

These potential biases are addressed by taking a more systematic approach, hence the name systematic reviews. Systematic reviews begin from a clearly articulated research question which is used to develop an explicit search strategy, with systematic screening of studies against explicit inclusion criteria. All studies found eligible for inclusion are systematically coded, with all findings relevant to the research question synthesized and reported. This is what makes a systematic review systematic: following a set process which seeks to minimize possible bias in the selection, analysis and reporting of research studies.

Systematic reviews were virtually unknown in international development outside of the health sector over ten years ago. But the last decade has seen an increase in the production of reviews, although sometimes the name is used without a full understanding of what the method implies. The principles of systematic reviews could usefully be applied more generally in evaluation. These principles are discussed briefly below.

A common complaint is that systematic reviews take too long for policy purposes. Policy makers need answers sooner rather than later and the 12-18 months to produce a review are too long to wait. There are three responses to this problem. The first is the need for a step change in the availability of evidence from systematic reviews in education, child welfare, agriculture, anti-poverty programs and so on, so that the evidence is there when needed. The Cochrane Library contains over 8,000 reviews on health. Campbell, which covers everything else, has just over 200. With the rise of impact evaluation in the last 15 years, there is huge potential for rapidly expanding the number of reviews, preferably produced under Campbell auspices for both coordination and quality purposes. A second response, is the need for more advance planning in the policy process. In health, WHO and national organizations like NICE in the UK plan what guidelines they will work on in the next cycle and so, after scoping, commission the necessary reviews to inform those guidelines. They allow that guidelines may take over two years to prepare because of the time needed for the reviews. Outside of health this approach is used in the Nordic model for the production of policy relevant reviews, with reviews taking one year to produce. The final response is to adopt rapid evidence approaches, though there are known biases in these approaches, so – despite their popularity – they should be seen as a last resort.

Implications for evaluation

First, a review begins with a clear definition of the scope, as captured in what is called the PICOS (Population, Intervention, Comparison or Context, Outcomes and Study design or Setting). It is useful to be clear on all of these elements ex ante, which applies to evaluations just as it does systematic reviews. In particular a clear focus on the intervention will help make for a more consistent approach and more straightforward interpretation of study findings. Evaluations are often weak on describing the intervention, and impact evaluations often ignore key issues in implementation and context.

The PICOS are part of the study protocol, which is an ex ante statement of the study design, which includes the PICOS, the data to be collected and how they will be found, and the data analysis plan. The search strategy lays out explicit inclusion and exclusion criteria for studies to be eligible for inclusion, thus removing bias, and possible accusations of bias, from selective inclusion. Analogously, evaluations should be clear on what data sources they should consider and why. Reviews of studies – especially qualitative studies including process evaluations – find they are often weak on describing their methods, including justification of data collection.

An equally important part of the protocol is the coding forms to be used for data extraction. These forms – like any data collection forms for a survey – need to be properly piloted. Sector and thematic evaluations will often look at a portfolio of projects should adopt the same systematic approach to collecting data from these projects. The testing of coding forms should be done by the coding team also as part of the training process. One thing we have learnt from both the replication crisis and research on research is that the research process is prone to errors. Systematic review approaches, such as proper testing of coding forms and the use of two coders, helps reduce these errors. Even with experienced coders, agreement rates on coding will typically be 80-85% (and will be 50-60% with less experienced coders). These discrepancies arise because of errors, ambiguities in the coding form and the need for judgement in data extraction. Good coding form design, testing and training can reduce all of these but never fully eliminate them. Again, these lessons apply to the data collection instruments used in evaluations – coding forms, interview schedules and so on – as they do to reviews.

It is not possible to completely remove judgement, which will have an important part in the evaluation process. But it should be understood what is meant by judgement. The use of judgment refers to the way in which the researcher or evaluator assigns a code based on the data to hand and the definitions provided of the coding category.

One part of coding is critical appraisal, that is assessing how much confidence can be based in study findings. To the extent that evaluations use different sources of evidence, they should be assessing the quality of that evidence, and couching their findings in language reflecting the confidence which can be had in those findings. A related issue for evaluations relying on stakeholder and key informant interviews is that of social relativity – a person's perspective on an intervention depends on their position in the process. Hence their statements may appear conflicting. The evaluator's role is not to identify which is correct and which is not, but together a whole picture consistent with these snapshots taken from different perspectives.

The approach is systematic in analysis and reporting, and thus avoids cherry picking which has been common in much evaluation research. One aspect of systematic analysis is reliance on meta-analysis for statistical analysis of effects rather than vote counting. Meta-analysis takes account of both sample size and the size of the effect which vote counting does not. Meta-analysis is not only about the average effect size but examining the causes of variations in effect, which may come from differences in context (including population characteristics), intervention design or implementation fidelity. There are several implications for evaluations here. For impact evaluations, should focus firmly on effect size rather than statistical significance. Second, evaluations need pay attention to heterogeneity. This point applies as much to qualitative studies – which are prone to cherry pick positive findings – as it does quantitative. Finally, meta-analysis weights evidence by how precise is the estimate of impact. Evaluations can also weigh up evidence by how representative it is likely to be of the experience of the population of interest.

Fourth wave: knowledge brokering and translation

As evidence has become more readily available as a result of the second and third waves of the evidence revolution so there has been increasing attention to the challenge of getting evidence. Researchers do not necessarily have the skills to ensure policy uptake of the findings of their research, neither do they face the incentives to do so. Knowledge brokering, or knowledge translation, has risen in response to this situation. Knowledge brokering is the process of rendering research findings into a form which is accessible and understandable to the intended users.

Knowledge brokering is the latest stage of the evidence revolution. There are two key aspects to the knowledge brokering movement. The first is recognition of knowledge brokering as a task in its own right: agencies have been set up with the purpose of knowledge brokering, and knowledge brokering is emerging as a profession.[11] The second key aspect are innovative

[11] A distinction can be made between knowledge brokering and knowledge translation. The former is rendering relevant and accessible presentations of the evidence, and the latter interpreting the

means of displaying evidence (research findings), going beyond good infographics (though there is nothing wrong with them) to interactive online evidence portals such as the Education Endowment Foundation's Teaching and Learning Toolkit.[12]

There are criticisms of knowledge brokering for oversimplifying the evidence – ignoring context, focus on a small number of quantitative outcomes and sidestepping complexity. But the more significant criticism is the underinvestment in knowledge brokering. Underinvestment in the whole evidence architecture is a problem in most sectors. You cannot build the top of the evidence pyramid without the lower layers.

Lessons for the evaluation community

As discussed above, evaluations can draw on systematic review methods. Evaluation agencies can also use evidence maps to identify what studies they have done, and when, and so identify gaps in their evidence base. This analysis of the supply side can be complemented by a demand side analysis from an evidence needs assessment.

Agencies which have made substantial investments in impact evaluation – such as Millennium Challenge Corporation (MCC), World Bank, Inter-American Development Bank (IADB) and Oxfam GB – made deliberate decisions to shift the focus of their impact evaluation programmes from accountability to lesson learning, and so making a more strategic selection of which evaluations to undertake. This approach mirrors the way in which the more advanced sections of the evidence community operate. For example, NICE in the UK identifies the systematic reviews it needs to inform guidelines to be published two years in the future.

There is a great need for evidence products to inform decision-making: portals, guidelines and checklists. Evaluation agencies can align their production with their agency's evidence needs. Some part of the evaluation department should be responsible for knowledge brokering of the agency's output – which, as stated above, is far more than one page summaries and dissemination events. Some part of the agency should be responsible for guideline development, with the study authors being engaged in that process.[13]

Evaluation agencies should invest in evidence products to support use of their evaluation findings. Some of these will be study-specific,[14] and others will support the use of evidence as it has accumulated over time. These approaches can help overcome inadequate institutional memory in development agencies, and the faddism to which they are prone.

meaning of evidence to take account of local context. In summarizing an evidence base for use by policy makers and practitioners both are likely to be necessary.

[12] https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/

[13] The production of evidence-based guidelines is a well-developed process. The Guidelines International Network is a network of agencies engaged in guideline production. The agencies subscribe to the use of GRADE which provides a set of standards for summarizing bodies of evidence for reviews and guidelines.

[14] BMJ Rapid Recs provide one possible model, and its further adaptation into MagicApp (https://app.magicapp.org/).

The fifth wave? Big data, AI and machine learning

Machine learning is being used in both evaluations and reviews. The use of machine learning in evaluations is still being explored – it can help identify patterns in the data (heterogeneity in impact) without imposing the researchers' priors - though may be biased by the results from which the machine learns, or other parameters given to direct it. Machine learning can thus be used to undertake textual analysis for evaluative purposes, and can also be used in quantitative impact evaluations.

Similarly, machine learning can assist systematic reviews in the screening process of eligible papers. Machine learning can also be used for reading and coding text and extracting data from tables, from which it is a short step to automated text and statistical analysis – hence automated living reviews and maps. There will be steps in this process requiring judgement so that needs to be allowed for to avoid error. But AI is already being used to produce news and financial reports. So, it should be possible to apply these same methods to summarizing studies. In principle, AI may avoid the cognitive biases humans have which can bias evaluations. Having said that, machines can only be as smart as the people they learn from and thus can reproduce their biases.

Lessons for evaluation

The main lesson is for evaluators to stay abreast of technological developments and to utilize them as appropriate. It will take time to learn from these experiences and so feed them through to standard evaluation practice. The technology now exists to make living evidence maps and reviews. This same technology could be applied to sector and thematic evaluations. In principle, an evaluation department could have a set of living sector and thematic reviews which are regularly updated as new source documents become available.

This would work by using machine learning both to identify relevant documents and also pick out key themes. Regarding the first item, sector and thematic evaluations often run into problems since they define a scope which does not readily correspond to an administrative category used by their agency. So, there is not clear universe for the evaluation to evaluate. This is a task ideally suited to machine learning. If the evaluation team can identify 20-30 relevant documents then they can let the machine learning algorithm lose on the agency's document repository to find similar documents. Hence, at very least the department. can use machine learning approaches to identify projects (or project components for intervention studies) and so maintain universes of studies for different sectors, themes or intervention categories.

But it is possible to go further than that. Semi-automated coding tools, such as Atlas-ti and NVivo, have been available for some years. These software can read texts, and create texts from audio recordings of interviews, focus groups etc, and extract key themes. As the software gets better, then the full process of document location and identifying key theme can be automated.

As a currently separate issue, for evaluation agencies conducting ex post study the increasing availability of Big Data increases the scope for these agencies to conduct rigorous impact evaluation. Such studies are already being done, such as Dolan et al's (2019) analysis of the impact of an insecticide treated bednet campaign on child mortality in the Democratic Republic of Congo. The CEDIL map produced by 3ie includes 437 studies which use Big Data to evaluate development outcomes (Rathinam et al., 2020).

Clearly these two approaches may merge, as evaluation machine learning algorithms process publicly available data to identify and answer evaluation questions.

Final words

There has been an evidence revolution over the last three decades. However, the evaluation community has not been a central player in this revolution. The first wave of outcome reporting was poorly linked to existing monitoring efforts. The second wave of impact evaluations has been undertaken by researchers rather than evaluators, whose main involvement has been to try to fight back the tide. The third wave of evidence synthesis has also largely engaged researchers despite the close overlap in approaches between many evaluations and the approaches used in reviews. And evaluators can learn from innovative approaches to the presentation of research findings. The same applies to the emerging fifth wave of new technologies.

The summary in Table 1 pulls out some of the main messages with respect to advancing the evidence revolution and the implications of each wave for evaluation.

Table 1 Summary of main points

Wave	Challenges and next steps	Implication for monitoring and evaluation
First: outcome monitoring	Develop and use core indicators for each sector	Move monitoring back from 'results' to feedback on project activities Evaluation departments conduct periodic assessment of agency M&E

Second: impact evaluation	Ensure impact evaluations answer questions of interest to decision-makers Do not use single studies to draw global conclusions	Undertake intervention-oriented evaluations rather than sector or thematic reviews Pay explicit attention to methods, reducing sources of bias in small n evaluations
Third: systematic reviews	More investment in systematic reviews is needed	Evaluations can adopt aspects of the systematic review approach to reduce possible biases
Fourth: knowledge brokering	Development agencies are yet to adopt evidence-based decision-making products based on systematic reviews	Develop knowledge products to increase use of evaluation findings
Fifth: AI and Big Data	Automating production of maps and reviews	Incorporate AI and Big Data into evaluation

References

Aiken, Alexander M., et al. "Changing use of surgical antibiotic prophylaxis in Thika Hospital, Kenya: a quality improvement intervention with an interrupted time series design." *PLoS One* 8.11 (2013): e78942.

Alkema, Leontine, et al. "Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group." *The lancet* 387.10017 (2016): 462-474.

Bennett, Andrew and Jeffrey Checkel (2014) *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge Universty Press.

Beach, Derek and Rasmus Brun Pedersen (2013) *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor: University of Michigan Press.

Black, Richard and Howard White (eds) (2002). *Targeting Development: critical perspectives on the Millennium Development Goals.* London: Routledge.

Blatter, Joachim, and Markus Haverland. *Designing case studies: Explanatory approaches in small-N research*. Palgrave Macmillan, 2012.

Byrne, D. and C. Ragin (2013) *The SAGE Handbook of Case-Based Methods*. London: Sage Publications.

Cartwright, Nancy. "A philosopher's view of the long road from RCTs to effectiveness." *The Lancet* 377.9775 (2011): 1400-1401.

Carvalho, Soniya and Howard White (1996) Indicators for Measuring Poverty Reduction. *World Bank Discussion Paper 254*. Washington D.C.: World Bank.

Clemens, M. A., and G. Demombynes. 2011. “When Does Rigorous Impact Evaluation Make a Difference? The Case of the Millennium Villages.” Journal of Development Effectiveness 3: 305–339. doi:10.1080/19439342.2011.587017

Copestake, J. Morsink, M. and Remant, F. (2019) Attributing Development Impact: The Qualitative Impact Protocol (QuIP) Case Book. Rugby: Practical Action Publishing.

Davies, Rick. "Representing theories of change: Technical challenges with evaluation consequences." *Journal of Development Effectiveness* 10.4 (2018): 438-461.

Dolan et al (2019) The impact of an insecticide treated bednet campaign on all-cause child mortality: a geospatial impact evaluation from the Democratic Republic of Congo, PLoS ONE 14(2)

Fretheim, Atle, and Oliver Tomic. "Statistical process control and interrupted time series: a golden opportunity for impact evaluation in quality improvement." *BMJ Quality & Safety* 24.12 (2015): 748-752.

Goetz, Gary and James Mahoney (2012) *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*, Princeton: Princeton University Press.

IEG (2013) World Bank Group Impact Evaluations: Relevance and Effectiveness. Washington, DC: World Bank. doi: 10.1596/978-0-8213-9717-6. License: Creative Commons Attribution CC BY 3.0.

Kusek, Zall and Ray Rist (2004) *Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners*. Washington D.C.: World Bank.

Leeuw, Frans and Jos Vaessen (2009) Impact Evaluations and Development Nonie Guidance on Impact Evaluation. Washington D.C.: IEG. World Bank.

Levine, Ruth and William Savedoff (2006) When Will We Ever Learn: using impact evaluation to improve development outcomes. *CGD Discussion Paper*. Washington D.C.: Centre for Global Development.

Mangam, Keith, et al. "Feasibility and Effectiveness of mHealth for mobilizing households for indoor residual spraying to prevent Malaria: a case study in Mali." *Global Health: Science and Practice* 4.2 (2016): 222-237.

Manning Richard and Howard White (2014) Measuring results in development: the role of impact evaluation in agency-wide performance measurement systems, Journal of Development Effectiveness, 6:4, 337-349, DOI: 10.1080/19439342.2014.989673

Masset, E. García-Hombrados, J. and Acharya (2020) Aiming high and falling low: The SADA-Northern Ghana Millennium Village Project, *Journal of Development Economics*, v143. https://doi.org/10.1016/j.jdeveco.2019.102427.

Mitchell, Shira, et al. "The Millennium Villages Project: a retrospective, observational, endline evaluation." *The Lancet Global Health* 6.5 (2018): e500-e513.

Natsios, Andrew (2010) The Clash of the Counter-bureaucracy and Development. Washington D.C.: Centre for Global Development. https://www.cgdev.org/sites/default/files/1424271_file_Natsios_Counterbureaucracy.pdf

Newman, John; Rawlings, Laura; Gertler, Paul; Newman, John*Rawlings, Laura*Gertler, Paul. 1994. "Using randomized control designs in evaluating social sector programs in developing countries (English)". *The World Bank Research Observer* 9(2): 181-201. http://documents.worldbank.org/curated/en/228191468774882721/Using-randomized-control-designs-in-evaluating-social-sector-programs-in-developing-countries

Picciotto, Robert (2014) 'Is Impact Evaluation Evaluation?' *European Journal of Development Research* 26(1): 31-38.

Ragin, Charles (2000) *Fuzzy Set Social Science*. Chicago: Chicago University Press.

Rathinam, F., Khatua, S., Siddiqui, Z., Malik, M., Duggal, P., Watson, S, and Vollenweider, X. 2020. Using big data for evaluating development outcomes: a systematic map. CEDIL Methods Working Paper 2. Oxford: Centre of Excellence for Development Impact and Learning (CEDIL). Available at: https://doi.org/10.51744/CMWP2

Stern, Elliot, Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, Barbara Befani (2012) broadening the range of designs and methods for impact evaluations DFID Working Paper 3. London: Department for International Development. https://www.oecd.org/derec/50399683.pdf

White, H. (2021) 'Using causal chain analysis in systematic reviews', *CEDIL Methods Brief 4*, London and Oxford: CEDIL.

White, Howard (2010) Contributions to Current Debates in Impact Evaluation. *Evaluation* 16(2):153-164.

White, Howard, and Edoardo Masset. "The Bangladesh integrated nutrition program: Findings from an impact evaluation." *Journal of international development* 19 (2007): 627-52.

White, Howard, and Daniel Phillips. "Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework." *New Delhi: International Initiative for Impact Evaluation* (2012).

White, Howard. "The twenty-first century experimenting society: the four waves of the evidence revolution." *Palgrave Communications* 5.1 (2019): 1-7.

Centre of
Excellence
for Development
Impact and Learning

www.cedilprogramme.org

@CEDILProgramme

Centre of Excellence for Development Impact