Introduction to this special issue

The rise of impact evaluations and challenges which CEDIL is to address

Evaluation plays a central role in the use of evidence. Evidence from evaluations about which programmes work, for whom, why and in what context is important for evidence-informed policy or practice.

The use of evidence in international development has developed through four waves (White, 2018b). The first wave was a focus on results. New Public Management in the 1990s was associated with the 'Results Agenda'. In international development, the 1992 'Wapenhans report' for the World Bank proposed that each project have a set of outcome-level Key Performance Indicators (KPIs). During the 1990s several development agencies adopted results frameworks, starting with USAID in response to the 1993 Government Results and Performance Act, and the UK Department for International Development (DFID) as part of the UK government's Modernizing Government agenda (GOUK, 1999). Globally, the adoption of the International Development Targets, later succeeded by the Millennium Development Goals (MDGs), became a global results framework for assessing development progress.

Before the results agenda, programme success was measured at best by activities or outputs, and often just by how much of the money had been spent. But national or global level indicators cannot tell us the role of any particular programme in achieving those results. Outcome monitoring – which is a before versus after analysis – cannot tell us if our programmes are working or not since many other factors are affecting those outcomes. There was a growing realization of this problem in the early 2000s. A letter from the US General Accounting Office (GAO) to USAID commenting on the 1999 performance report stated that the goals were: 'so broad and progress affected by many factors other than USAID programmes, [that] the indicators cannot realistically serve as measures of the agency's specific efforts' (GAO, 2000). One of us wrote a paper for the UK National Audit Office review of DFID's performance measurement making the same point.[1]

The second wave of the evidence revolution was the rise in impact evaluations of development interventions. Newman et al. (1994) laid out the case for rigorous impact evaluation in international development in an early paper in the *World Bank Research Observer* as the World Bank supported early RCTs in Latin America. It was nearly ten years later until the J-PAL and the World Bank's Development Impact Evaluation initiative (DIME) were set up in 2003 and 2004 respectively. In 2006 the Centre for Global Development published the influential report *When Will We Ever Learn?* (CGD, 2006) arguing the case for 'a global fund for impact evaluation' leading to the creation of the International Initiative for Impact Evaluation (3ie).

As has been documented in two papers previously published in this journal (Cameron et al., 2016 and Sabet and Brown, 2018), there has been a nearly exponential growth in impact evaluations over the last 15 years – especially randomized controlled trials. In 2000 there were around 50 impact evaluations

[1] The NAO report is NAO (2002). More detail can be found in White (2005)

published, nearly all in health. By 2011, over 500 studies were published in a single year in different sectors of interventions such as agriculture, education, nutrition and governance.

As the literature has grown it has given rise to a new challenge: how to stay on top of a growing body of evidence, and how to reliably inform policy in an unbiased way. Policy should not be informed by single studies because single studies are rarely able to provide recommendations that apply to a plurality of contexts and of characteristics of implementation. For example, farmer fields school training seems to work in reducing pesticide use and increasing yields, but only in pilots, while programmes at scale show no effects (Waddington et al., 2014). School management seems to improve school outcomes, but not in low income settings (Carr-Hill et al., 2018). Land titling improves productivity, but not in Africa (Lawry et al., 2014). One example of the difficulty to extrapolate results from single studies is deworming. One study has carried a lot of weight in policy discussions (Miguel and Kremer, 2004), leading deworming to be seen as a best buy in development. We are not debating here the merits of that study, which were many. But other studies of the same intervention produced different results. How should a policy-maker deal with a body of conflicting evidence? The answer is that systematic reviews offer the best, least biased, approach to summarizing bodies of evidence.

The growth of systematic reviews is the third wave of the evidence revolution. When 3ie issued its first Request for Proposals for systematic reviews in September 2008 they were virtually unknown in international development outside the health field. Fewer than 20 reviews were published in 2008. By 2016 over 100 were published that year alone (source: 3ie database).

Returning to deworming, there are 65 studies of deworming covered in the Campbell review of the intervention (Welch et al., 2016), nearly all of which find no impact on health, nutrition or education outcomes. As a result, average effect sizes in the meta-analysis are not significantly different from zero.[2] There is at very least a puzzle here to understand in which circumstances deworming is a priority programme, or where, whilst desirable, it is not the best use of scarce resources.[3]

The fourth wave of the evidence revolution is knowledge translation or knowledge brokering. It is not enough to publish high-quality evidence syntheses. Translating that evidence to inform policy or practice is a separate activity which goes far beyond traditional dissemination. In response to the global evidence, WHO has revised their guidelines to recognize the importance of improved water and sanitation in achieving sustained reductions in the worm burden and the consequent health benefits (WHO, 2017). This step reflects WHO's institutionalization of the use of evidence from systematic reviews: WHO's guidelines on producing guidelines state that such

[2] Reweighting the effect size estimates to get a significant effect, as has been done by some critics, is to miss the point. There are three studies showing substantial effects, all from sub-Saharan Africa. In the face of such heterogeneity the role of meta-analysis – or any good statistical analysis – is to analyze the sources of variation not focus on the average effect. Unfortunately, we do not yet know the factors underlying the 'African exceptionalism'.

[3] This is not to argue the case for screening. The cost of screening is undoubtedly far higher than the cost of universal treatment. It is rather to make the case for area targeting based on geographical proxies for effectiveness.

guidelines have to be based on high quality systematic reviews (WHO, 2010). Outside of health, the *What Works* movement in the UK are building evidence platforms to make evidence available to decision-makers – and in the best cases that evidence is based on systematic reviews (Gough and White, 2018).

So far, so good. We have both been involved in various stages of this evidence revolution and we must surely be counted amongst 'the believers'. But there are limitations. One is that, whilst RCTs have been applicable in a far broader range of contexts than was originally imagined even by their proponents, there are important questions of effectiveness which are not amenable to large n impact evaluations. In establishing CEDIL, DFID challenged us to answer difficult questions such as 'what have our package of interventions over the last ten years done to improve women's empowerment in Afghanistan?' or 'what is the effectiveness of budget support?'

A second issue is that impact evaluations, systematic reviews and What Works evidence platforms are primarily focused on the impacts of interventions and their effect sizes[4]. But we know implementation matters. There has been too little attention on second generation questions of design and implementation in both impact evaluations and reviews,[5] although these are often the questions of most interest to decision-makers. Studies addressing such questions are very likely to require mixed methods designs which effectively integrate the two approaches. But successful mixed-methods designs remain the exception not the norm.

A third, and final, issue is that policy uptake of research findings remains problematic. International development has no evidence platforms of the sort now common in the UK and US under the auspices of the *What Works* movement. Very little is known about what are the best methods to disseminate knowledge and to promote the assimilation and understanding of evidence. It was to address all these issues that DFID supported the creation of a new Centre of Excellence for Development Impact and Learning.

What is CEDIL?

CEDIL was established in 2017 by DFID with the goal of developing and supporting new evaluation methods in under-researched areas of international development. In particular, CEDIL was created to fill major gaps in: evidence, methods, synthesis, and translation. CEDIL was set up to achieve the following goals: developing new evaluation methods in unexplored thematic areas, commissioning impact evaluations and related research, and promoting evidence use.

CEDIL is composed of a Research Directorate, which provides strategic direction and technical guidance, and of a Programme Directorate, which manages the project. The Research Directorate includes a consortium of five institutions: the International Initiative for Impact Evaluation (3ie), the Campbell Collaboration, the Centre for Evaluation of the London School of Hygiene and Tropical Medicine, the

[4] The term effectiveness is used in many circles to refer to what impact evaluators call impact. However, other evaluators use the term impact to refer to top level indicators resulting in a differing meaning of the term impact evaluation. See White (2009) for more discussion.

[5] For a discussion of the neglect of second generation questions, see for example Waddington et al. (2018) for systematic reviews and Melvin and Lenz-Watson (2011) for experimental studies.

EDePO at University College London, and the EPPI-Centre. Internationally recognised experts in evaluation constitute the Intellectual Leadership Team of CEDIL and activities are overseen by an Advisory Board. A small Research Directorate is based at the London International Development Centre. The Programme Directorate is led by Oxford Policy Management (OPM).

The role of the Research Directorate is to identify programmes of work for CEDIL to address the evaluation challenges mentioned above. The CEDIL Programme Directorate will issue Requests for Proposals to commission impact evaluation studies, systematic reviews, and academic papers under these programmes of work.

To inform the formulation of the programmes of work, in its inception phase CEDIL produced nearly 20 papers addressing methodological issues in impact evaluation, synthesis of evidence, and knowledge translation. Some of the papers were think pieces, others were summaries of expert consultations, others still were methodological surveys in a particular area. All CEDIL papers are available on the programme website and summary versions of five of these papers are included in this special issue.

Each paper tackled one of the most debated topics in the impact evaluation community today such as, for example, the production of timely evaluations and systematic review, the generalisability and transferability of findings (external validity), and the successful involvement of stakeholders in the evaluation process. Despite the variety of methods and approaches discussed, three common characteristics across the papers stand out: an interdisciplinary approach to methodological development, a rejection of hierarchies of methods, and a preference to explaining results versus testing hypotheses.

CEDIL member institutions and experts come from different disciplinary backgrounds such as economics, epidemiology, education, political science and environmental science. The evaluation methods proposed in the papers went through rounds of interdisciplinary discussions and workshops. In some cases, the methods proposed by CEDIL are not absolutely new but are new for a particular discipline or area of application. It is hoped that cross-fertilisation across disciplines will lead to a better integration of methods and, perhaps, to the formulation of novel methodologies.

CEDIL is committed to rigorous evaluation and scientific research but does not adhere to any methodological hierarchy. CEDIL members have different methodological inclinations but the Centre does not spouse any specific approach. Methods are valued for their ability to explain phenomena and to inform policies.

Many CEDIL papers reflect an interest in explaining observed phenomena as the basis for formulating sound policies. This approach aims at balancing the recent popularity of experimental approaches in development economics. CEDIL recognises the value of randomised control trials in the generation of credible evidence but promotes the embedding of experiments in middle-range theories and behavioural models to better inform policy and projects.

The papers in this issue

This special issue of the Journal of Development Effectiveness features a subset of the studies that were presented at the CEDIL conference held in London on 25th January 2018. We selected five papers for this issue following no particular criteria except for the need to allow one paper from each CEDIL partner institution and at the same time paying attention to the interests of the Journal's readership. The papers selected are reflections on methods used and proposed in impact evaluations and systematic review. All papers were peer-reviewed by two members of the CEDIL Intellectual Leadership Team and by DFID. In addition, each paper was reviewed by a blind referee for the Journal version.

As mentioned earlier, the collaborative nature of CEDIL resulted in cross-fertilisation of approaches and themes between authors of different disciplines. Some unifying themes have emerged that are common to several papers. In other cases, authors have pursued standalone methodological investigations. Common threads which emerged were the role of theory of change analysis and the use of mixed methods. These themes are reflected in the first three papers in this collection.

The papers by Kneale et al. and Davies are both concerned with theories of change. They both make the point that there are many types of causal relationship. As Davies points out, models typically do not mention the timing, duration or sequencing of relationships. And they are ill suited to capture the nature of the relationship. The authors of the two papers provide a list of common violations of causal linearity assumptions, such as threshold and plateau effects, tipping points, necessary or sufficient conditions, interactive effects and vicious or virtuous circles.

Kneale et al. argue that the diverse ranges of causal relationships require different approaches to assessing causality, and that systematic reviews are in principle well suited to this approach, although in practice many are bare bones reviews which do not look beyond effectiveness (Snilstveit, 2012). In particular analysis along the causal chain is likely to rely on different methodological approaches, thus requiring the use of mixed methods. As found by Jimenez et al. in their discussion of the use of mixed methods in reviews, integrated reviews go 'beyond the sum of their parts'. We return to the Jimenez et al. paper below, but here we note that both that paper and Kneale et al. advocate the explicit identification of a strategy for integrating mixed methods with a justification for the approach. Kneale et al. also propose that causal chain based reviews should update theories of change in the light of new evidence.

Much of the discussion by Kneale et al. is echoed by Davies who identifies the following issues with typical representations of the theory of change: (i) unlabelled connections, a simple arrow tells us nothing about the timing, duration, scale or type of causal connection, (ii) missing connections, especially if the theory of change is presented in silos (see also White, 2018a), which detracts from the usefulness in the theory of change in identifying key evaluation questions, (iii) symmetrical designs in which aesthetics triumph over utility; (iv) nested hierarchies and heterarchies which present complex theories of change with many arrows and clusters of causal effects, but no indication of timing, sequencing or relative importance, (v) feedback loops, whilst in principle theories of change can readily accommodate feedback loops

(reverse or simultaneous causation), in practice most do not – and if too many such loops are included then there may be no clearly describable causal pathway, (vi) wider connections which lead to over-optimistic theories of change which fail to anticipate external factors inhibiting programme effectiveness or to identify contextual factors which mediate it.

In response Davies identifies six solutions. Four of these concern using approaches which help identify the most important causal mechanisms to be tested: network analysis, participatory network analysis, sensitivity analysis (predictive modelling) and dynamic approaches. One is better specification of the type of causal relationship, as discussed above. Finally, Davies recommends using better software, a suggestion also made by Kneale et al.

Jimenez et al. review the use of mixed methods in 40 impact evaluations and seven systematic reviews. To do this, they present a tool which assesses the quality of both the quantitative and qualitative analysis in the studies as well as the integration of the two approaches. The 40 IEs are all what the authors call 'quantitatively-driven' studies, i.e. have a clear identification strategy at the heart of their research design, but which also explicitly mention at least one qualitative component to the research design. The tool scored both quantitative and qualitative designs on a scale from 0 to 130. The overall rigour of the quantitative designs was substantially higher than that of the qualitative designs, with average scores of 96 and 43 respectively. The average integration score was three out of a possible six. Analysis shows that studies with better qualitative designs were more likely to score highly on integration. Other factors associated with high scores on mixed methods were (i) an explicit rationale for integration and its value added for the study, (ii) a multidisciplinary team, and (iii) adequate documentation of approaches.

The final two papers are concerned with other issues flagged in the CEDIL inception papers: measurement and timely evaluation. The fourth paper, by Almas et al., was led by a team based at the Institute of Fiscal Studies, with the support of two international experts and CEDIL staff, and discusses measurement issues. Impact evaluations are designed to assess how specific interventions affect outcomes such as 'poverty', 'empowerment', 'child development' or 'environmental sustainability.' While these theoretical constructs are well defined, applied researchers can rarely measure them directly and have often to rely on proxy indicators. In addition, the measurement of the indicators is often difficult to perform and conducted with considerable error. This paper shows how these difficulties can be addressed: by formulating new measures altogether, by using multiple measures of the same construct, and by employing new machine learning methods for outcome classification.

In relation to the first point, the authors argue that new measures should be tailored to the specific behavioural models that the researchers want to estimate as they should aim at recovering the main parameters of such models. In relation to the second point, the authors argue that an index of multiple imperfect indicators may be preferable to a single perfect, but unattainable, indicator. Finally, the authors show how the availability of new and large datasets and of statistical methods of machine learning has opened up the possibility to measure new constructs with limited data collection effort. The paper also includes a useful survey of sources of available

datasets of secondary data that can be used in quasi-experimental studies at the design and analysis stage.

The fifth and final paper, by Jayne Webster et al., was led by a team based at the London School of Hygiene and Tropical Medicine and features experts in conducting agile and timely evaluations in international development. It is well known that high quality impact evaluations take a long time to complete – 3-5 years being the norm. This timeframe does not always help policy-making. Results of impact evaluations often come too late once decisions have already been made, and large impact evaluation studies offer limited support to improving programmes while they are being implemented. The problem is well-known in public health where decisions about programmes need to be made quickly with any information available, and where methods to quickly collect and analyse information have been produced and tested. The authors provide an overview of the best approaches and tools available for conducting timely evaluations, together with a critical appraisal and a framework that helps researches identifying which methods are best suited to achieve different evaluation goals. The review includes a description and an appraisal of, among others, behavioural centred designs, statistical process control, bottleneck analysis, A/B testing, adaptive trials, and qualitative impact assessment protocol. The authors highlight the absence of empirical studies and reflect on the need of testing and validating these tools in the field.

We hope that the papers published in this special issue will stimulate researchers to pursue impact evaluations and synthesis of evidence in new areas and to use new approaches and methods. Rigorous research should not occur to the detriment of imagination and creativity. The papers presented in this special issue are only a sample of the wide range of methodological and theoretical investigations explored by CEDIL. They are part of a larger intellectual effort that will guide our research agenda over the coming years.

Howard White and Edoardo Masset

References

Cameron, D. B., Mishra, A., and A. N. Brown (2016) The growth of impact evaluation for international development: how much have we learned?, *Journal of Development Effectiveness,* 8:1, 1-21, DOI: 10.1080/19439342.2015.1034156

Carr-Hill R, Rolleston C, Schendel R., and H. Waddington (2018) The effectiveness of school-based decision making in improving educational outcomes: a systematic review, *Journal of Development Effectiveness,* 10:1, 61-94

Centre for Global Development, CGD (2016) 'When Will We Ever Learn: Improving Lives through Impact Evaluation', Washington DC: CGD.

General Accounting Office (2000) Observations on the US Agency for International Development's Fiscal Year 1999 Performance Report and Fiscal Years 2000 and 2001, Performance Plans. Washington DC: GAO.

Government of the United Kingdom (GOUK) (1999) White Paper: Modernising Government. London: GOUK.

Gough, D. and H. White (2018) 'Evidence standards and evidence claims in web based research portals', mimeo.

Lawry, S, Samii, C, Hall, R, Leopold, A, Hornby, D, Mtero, F. (2017) The impact of land property rights interventions on investment and agricultural productivity in developing countries: a systematic review. *Journal of Development Effectiveness,* 9:1, 61-81.

Melvin, M. M. and A. L. Lenz-Watson (2011) "Ethics and the conduct of randomized experiments and quasi-experiments in field settings" in Printer, A.T. and S. K. Sterba, *Handbook of ethics in quantitative methodology*. Routledge, Avignon, Oxon.

Miguel, E. and M. Kremer (2004) Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, *Econometrica*, Vol. 72, No. 1, 159–217

Newman, Rawlings, L. and P. Gertler (1994) Using Randomized Control Designs in Evaluating Social Sector Programs in Developing Countries, *The World Bank Research Observer*, Volume 9, Issue 2, 1 (July), Pages 181–201

National Audit Office (2002) Performance management – helping to reduce world poverty. London: The Stationery Office.

Sabet, S. M. and A. N. Brown (2018) Is impact evaluation still on the rise? The new trends in 2010–2015, *Journal of Development Effectiveness*, 10:3, 291-304, DOI: 10.1080/19439342.2018.1483414

Snilstveit, B. (2012) Systematic Reviews: from 'bare bones' reviews to policy relevance, *Journal of Development Effectiveness*, 4:3, 388:408.

Waddington, H, Snilstveit, B, Hombrados, J, Vojtkova, M, Phillips, D, Davies, P and White, H. (2014) Farmer Field Schools for Improving Farming Practices and Farmer Outcomes: A Systematic Review, Campbell Systematic Reviews 2014:6

Waddington, H., Masset, E. and E. Jimenez (2018) What have we learned after 10 years of systematic reviews in international development?, *Journal of Development Effectiveness*, 10:1, 1-16.

Welch V. A., et al. (2016) Deworming and adjuvant interventions for improving the developmental health and well-being of children in low- and middle-income countries: a systematic review and network meta-analysis, Campbell Systematic Reviews, 2016:7

White, H. (2005) The Road to Nowhere? Results-based management in international cooperation. In Cummings S. ed. *Why did the chicken cross the road? And other stories on development evaluation*. Amsterdam: Royal Tropical Institute (KTI).

White, H. (2009) 'A Contribution to Current Debates in Impact Evaluation' *Evaluation* 16(2): 153 – 164.

White, H. (2018a) Theory-based systematic reviews, *Journal of Development Effectiveness*, 10:1, 17-38, DOI: 10.1080/19439342.2018.1439078

White, H. (2018b) 'The four waves of the evidence revolution: the role of systematic reviews' Campbell Collaboration Discussion Paper.

World Bank (1992) 'Effective Implementation: Key to development impact' Washington D.C.: World Bank.

World Health Organization (WHO) (2010) WHO Handbook for Guideline Development. Geneva: WHO.

World Health Organization (WHO) (2017) Preventive chemotherapy to control soil-transmitted helminth infections in at-risk population groups: Guideline. Geneva: WHO.