palgrave
communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

COMMENT

https://doi.org/10.1057/s41599-019-0253-6 OPEN

The twenty-first century experimenting society: the four waves of the evidence revolution

Howard White[1]

ABSTRACT
This paper presents a personal perspective-drawing especially on the author's experience in international development—of the evidence revolution, which has unfolded in fours waves over the last 30 years: (1) the results agenda as part of New Public Management in the 1990s, (2) the rise of impact evaluations, notably randomized controlled trials (RCTs) since the early 2000s, (3) increased production of systematic reviews over the last ten years, and (4) moves to institutionalize the use of evidence through the emergence of knowledge brokering agencies, most notably the What Works movement in the United States and the United Kingdom. A fifth wave may come from the potential from AI, machine learning and Big Data. Each successive wave has built on the last, and together they comprise the supply side of the evidence architecture. To support the use of evidence demand side activities such as Evidence Needs Assessments and Use of Evidence Awards are proposed.

[1] The Campbell Collaboration, Delhi, India. Correspondence and requests for materials should be addressed to H.W. (email: hwhite@campbellcollaboration.org)

Introduction

Nearly two-thirds of schools in England use evidence from systematic reviews to decide how to spend school resources and plan classroom activities. The US development NGO, International Rescue Committee (IRC), has committed to making all its programmes evidence-based or evidence-generating by 2020. In December 2018 US Congress passed the Evidence-Based Policy Making Act. In the UK and the US the 'what works movement' provides evidence on the effectiveness of interventions to improve learning, reduce child abuse and homelessness, fight crime and improve well-being.

Have we achieved Donald Campbell's vision of the experimenting society?[1] That is, a society in which social policy choices are informed by evidence from high quality research–'testing by piecemeal social engineering' (cited by Campbell, 1988). Whilst such experimenting was occurring since the 1930s–mainly in the United States–there has been a step change in recent years partly enabled by the What Works movement. It is fair to speak of an evidence revolution. This revolution started in the health sector with evidence-based medicine going back seventy years (Oliver and Pearce, 2017). In other sectors, such as international development, education and social welfare, the evidence revolution has broadly followed the four waves described in this commentary. The following narrative describes most closely the experience in international development. The narrative is focused on what has been done, and can be done, to support the use of evidence in decision-making. Of course it is recognized that many other factors affect decision-making. Policy is ultimately a political process (see, for example, Cairney, 2016, and Parkhurst, 2017), but those issues are not further considered here.

There has been growing attention to use of evidence to inform policy, with recent reviews of what that literature tells us; e.g., Langer et al. (2016), and Oliver and Cairney (2019). However, the main focus of this literature is on the approaches researchers can take to support the use of research findings in policy (e.g., Evans and Cvitanoivc, 2018), such as engage users in the setting of research questions or the production of the research itself. This paper is as much concerned with the demand side as the supply side, describing initiatives from research commissioners and users, not just producers and how to support demand. In particular, a central focus is the institutionalization of the use of evidence.

This institutionalization can be seen in the four waves of the evidence revolution which are: (1) the results agenda, (2) impact evaluations, (3) systematic reviews, and (4) knowledge brokering (see Fig. 1). This paper describes the evolution of the revolution through these four waves with examples from around the world, but mostly from my own background of international development.

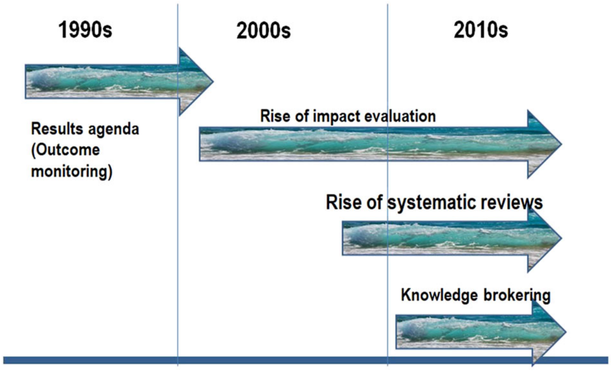

Fig. 1 Four waves of the evidence revolution

The first wave: the results agenda, 1990

The evidence revolution emerged as part of New Public Management which took hold in Anglophone and Scandinavian countries in the 1990s. Notable landmarks include the 1993 Government Performance and Results Act (GPRA) in the US and the 1999 Modernizing Government White Paper in the UK. New Public Management held government agencies to account for their performance as captured by trends high-level outcomes (results) such as unemployment, poverty, and so on. This shift to a focus on outcomes was an important achievement as performance had previously been assessed simply by inputs such as how much money had been spent.

One consequence of the focus on outcomes was an effort to establish better indicators. So in the mid90s the World Bank published a series of sector reports on preferred indicators. My colleague Soniya Carvalho and I authored 'Indicators for Monitoring Poverty Reduction' as part of that effort (Carvalho and White, 1994). As another example, the United States Agency for International Development (USAID) produced a 'Handbook of Democracy and Governance Program Indicators' (Center for Democracy and Governance, USAID, 1999). These efforts are worthwhile and should be revisited as the lack of consistency in measurement which persists in many sectors makes meaningful comparisons in performance between programmes difficult.

More generally in the international development domain the results agenda manifested itself in the International Development Targets (IDTs) which were replaced by the more widely-adopted Millennium Development Goals (MDGs), now succeeded by the Sustainable Development Goals (SDGs). 'Results frameworks' became common across development agencies.

All this was very laudable. There was just one drawback: 'results' don't measure agency performance.

Performance measures can be assessed against the triple A criteria of alignment, aggregation and attribution.[2] Alignment: are the measures aligned with the agency's goals? Outcome measures do well on this criterion. Aggregation: can the measures be aggregated across the agency to give a single figure for agency performance? Again, outcome measures do well. Attribution: can changes in the measure be attributed to the efforts of that agency? Here outcome measures fall down, as the case of USAID shows.

In response to GPRA, USAID started to publish Annual Performance Reports showing results against their strategic goals, such as growth rates in the main recipients of US foreign assistance. In their review of the 2000 performance report the General Accounting Office (GAO) wrote to USAID 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' (General Accounting Office, 2000). In response USAID abandoned using indicators related to the strategic goals ('results') to measure USAID's performance.

My own engagement with these initiatives arose when the UK National Audit Office asked me to undertake an assessment of DFID's performance measurement system as background for their own report (White, 2002, and NAO, 2002, respectively), which concluded that 'one must wonder on what data DFID management do base their decisions. There is no "bottom up" system to indicate overall performance. And the IDT-related indicators embodied in the PSA [the DFID results framework] are of little operational use'. A short summary entitled 'Road to Nowhere' warned that USAID had been down the results road, but they had come back saying there was nothing down there (White, 2005b). Unfortunately that call was not heeded, and many agencies and developing country governments embraced results frameworks, and some continue to do so.

But there was something else happening also. My paper for NAO argued for the use of logic models (or theory of change) to tackle the attribution issue. But there is another way: impact evaluations which measure what difference an intervention makes. There were some such studies already, but the number of published impact evaluations began to grow rapidly in the first decade of this century. Particularly prominent and contentious was the use of randomized controlled trials (RCTs). This was the second wave of the evidence revolution: the rise of RCTs.

The second wave: the rise of RCTs approx. 2003

RCTs of social programmes were not new. They have been carried out, mainly in the United States, since the 1930s (Oakley, 1998). But across all sectors and across the world, there is a clear upward trend in the number of RCTs, and other impact evaluation designs, being published from the early 2000s.

In international development there had been a few RCTS of interventions in the 1990s–most famously the Progressa conditional cash transfer programme in Mexico. But the movement took off in the early 2000s. Two prominent organizations supporting development RCTs–J-PAL and IPA–were founded in 2003 and 2005 respectively. More significant was the institutionalization of impact evaluation under the Development Impact Evaluation programme, DIME, at the World Bank in 2004 which provided seed finance to support the design of World Bank funded interventions. The Washington-based think tank, the Centre for Global Development (CGD), issued the influential report *When Will We Ever Learn?* berating the development community for spending billions of dollars on programmes for which there was no evidence (Levine and Savedoff, 2006). The CGD campaign mobilized bilateral agencies and philanthropic foundations to support the creation of the International Initiative for Impact Evaluation (3ie) in 2008. These efforts led to a substantial increase in the production of impact evaluations of international development interventions, a trend which was mirrored in other sectors.

I moved to the World Bank in 2002 to lead a small programme of impact evaluations in the Independent Evaluation Group, leaving in 2008 to be the founding Executive Director of 3ie. At the World Bank I had led four studies, but during my time at 3ie we funded close to 200. One of the first things we did at 3ie was to start a database of development impact evaluations. That database now contains close to 5000 studies impact evaluations. There were fewer than 50 impact evaluations a year being published in 2003 rising to over 500 a year by 2012.

As mentioned above, similar trends can be seen in other sectors; though the timeline health predates this. In education around 10 RCTs were being published each year in the early 2000s, growing from 2003 to over 100 a year by 2012 (Connolly et al., 2018). For social work the numbers are around 10 RCTs a year in the early 2000s and over 50 by 2012 (Thyer, 2015).

The findings from this blossoming of impact evaluations have shown the importance of conducting such studies. It appears that there is in general an 80% rule. That is 80% of things don't work. In education, 90 interventions evaluated in RCTs by IES—90% had weak or no positive effects. For employment and training programmes 75% of RCTs commissioned by the Department of Labor show weak or no positive effects. And in the private sector, over 13,000 RCTs of new products/strategies conducted by Google and Microsoft report no significant effects in over 80% of cases (Pfeffer and Sutton, 2006). A study by the European Commission found that 85% of projects financed under the Clean Development Mechanism were actually unlikely to provide additional reductions in carbon emissions (Cames et al., 2016). The effective altruism Oxford-based NGO, 80,000 h concluded that 80% was probably a generous figure—more likely a higher percentage of things don't work (Todd B and the 80,000 hours team, 2017). So, as good Bayesians, in the absence of evidence to the contrary from a rigorous evaluation, we should assume our programme doesn't work.

The more evidence-oriented development agencies—such as the UK Department for International Development and the Bill and Melinda Gates Foundation—require a statement of evidence from rigorous studies to support new proposals, and, in the case of DFID, how the proposed activity will collect the needed evidence if it doesn't exist.

Since so many things don't work, rigorous evaluation is great value for money. The evaluation of Mexico's conditional cash transfer programme, Progesa, in the mid-90s cost US$2 million. The evaluation found strong effects on education, health, nutrition and poverty, generating political support for the programme so it survived political transitions. Generously assuming that without the evaluation funds would have otherwise been used on a programme which was half as effective, the use of the evaluation findings resulted in an additional 550,000 children making the transition to secondary school and 800,000 children aged 12 to 36 months having reduced stunting in the years between 2000 and 2006.[3]

With so many studies it becomes hard to stay on top of the literature. Decision-makers anyway are unlikely to read academic papers, but may be influenced by findings from high profile studies. But decision-making should be based on an assessment of the body of evidence, not single studies. I take one, admittedly contentious, example to illustrate this point: school-based deworming.

An influential study from Kenya shows strong effects of deworming on nutrition, health and education outcomes (Miguel and Kremer, 2004). This study in particular has influenced the Deworm the World movement. But, as reported in Cochrane (Taylor-Robinson et al., 2015) and Campbell (Welch et al., 2016) systematic reviews, the vast majority of studies show no such effects. There is a puzzle to explain the African exceptionalism, and understanding that would help design and target programmes in a cost effective manner. But for most the world it seems that it is not so the deworming is 'the best buy in development' as some claim–we should not be misled by single studies or a small number of studies when there is a larger body of literature.

The third wave: the rise of systematic reviews 2008

This need to draw on bodies of evidence has powered the third wave of the evidence revolution: the rise of systematic reviews. In most sectors this wave has taken place over the last ten years. This wave came earlier for health, laying the basis of Evidence-Based Medicine, driven by the Cochrane Collaboration and World Health Organization (WHO). Other sectors have followed more recently.

Again, this wave is across countries and sectors. In social policy there were few systematic reviews published before 2000, around 25 a year in the noughties, growing from 2010 to 230 published in 2016.[4] In international development there were few reviews before 2008, after which the number grew steadily to over 100 published in 2016.[5] In education a few reviews were published a year in the early 2000s, rising to the 20 s toward the end of the decade and over 200 in 2018.[6] This increase has been driven in part by the What Works movement which I discuss in the next section.

My organization, 3ie, played a role in the rise of reviews in International Development. We issued a call for proposals for nearly 60 reviews in 2010, and have managed the funding for over 100 reviews. We partnered with the Campbell Collaboration–the international research network promoting the production and use of high quality systematic reviews—in 3ie's work on systematic reviews. In 2010 we set up the Campbell International Development Coordinating Group which is housed in 3ie's London office.

Some reviews support the rather pessimistic view of programme effectiveness. The first Campbell review published showed that Scared Straight programmes actually make youth more likely to become criminals rather than less (Petrosino et al., 2013). A review of teenage pregnancy programmes found none to be effective in reducing sexual activity or pregnancy (Scher et al., 2006).

I left 3ie in 2015, taking up the position as CEO of Campbell toward the end of that year. Supporting the production of reviews is Campbell's core business. A first step was to put in place a new strategy with two key goals: more reviews, and more use of reviews.

The goal of more reviews is being pursued in various ways. One important way is our combined training and mentoring for new research teams, especially in low- and middle-income countries, which has resulted in a step increase in Campbell Library publications. This approach is paying off. We published 103 papers in the Campbell Library in 2018: double the number published three years' earlier in 2015.

The challenge in promoting the use of systematic reviews is that they are long, technical documents. They also may not be accessible either in terms of discoverability and accessibility—hard to find or behind a pay wall—or in terms of comprehensibility. A broad review may well run into several hundreds of pages. And the implications for policy may not always be clear. Getting review findings into policy and practice has been the fourth wave of the evidence revolution: knowledge brokering or knowledge translation.

The fourth wave: the rise of knowledge brokering 2010

Activities in the fourth wave seek to institutionalize the use of evidence in policy and practice. There are two ways of doing this: direct interaction—which I call the Nordic model—and creating knowledge products such as evidence portals—which is that What Works movement. Whilst some of these initiatives predate the current decade it is this decade which has seen What Works gain the momentum to be called a movement.

Each of Denmark, Norway and Sweden has 'knowledge centres' for education, health and social welfare. These are government-funded research centres. Government-funded research centres are not unusual. What is different about the Nordic model is that they have staff whose regular job is producing reviews for to inform government decision-making. These are not academic researchers whose incentives are to publish. They are researchers whose incentive is to produce systematic reviews relevant for policy and practice. The research teams meet regularly with government agencies to agree priority topics, and to discuss emerging findings and how they should be interpreted for policy purposes. This model is also commonly adapted by teams which provide rapid evidence responses rather than full systematic reviews of which there are a growing number.

The direct interaction model can work when dealing with a small number of decision makers, say in central government or in a single agency. It is less well suited when decision-making is decentralized to district school superintendents or head teachers, or by prison governors, or by social work teams, or by one of the many thousands of development NGOs. In these cases evidence products which can be used by decision-makers without support required.

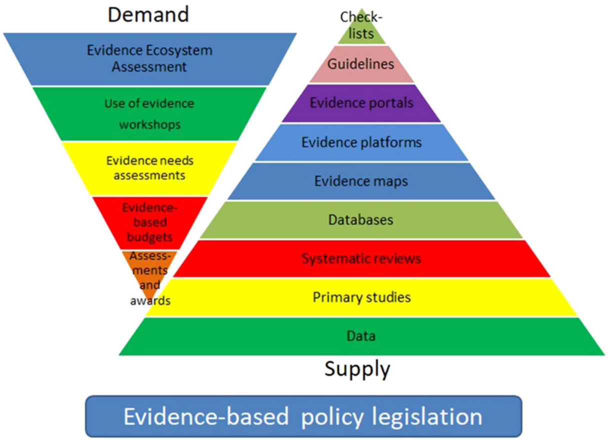

Fig. 2 The evidence architecture

But the approach is spreading. I see this as the true manifestation of the fourth wave: building the evidence architecture to institutionalise the use of evidence. This architecture is shown in Fig. 2. Institutionalisation can be underpinned by legislation requiring evidence-based policy, as passed in the United States in December 2018 or Mexico's 2004 Social Development Law which required external evaluation of all government-funded social programmes. Such legislation requires government-funded agencies to produce and use rigorous evaluations. A description of the various levels of the pyramid follows, starting on the supply side.

The layers of the supply-side pyramid do not represent standards of evidence as in the conventional evidence pyramid. Rather they reflect high degrees of knowledge translation and curation. Hence data are analysed and summarised in studies. Those studies are in turn analysed and summarized in systematic reviews.

Databases contain studies and reviews related to a specific sector and possible specific research designs. There are many such databases: the US Institute of Education Sciences' (IES) ERIC database for education research, Epistimonikos for systematic reviews and impact evaluations in health, the Global Policing Database for interventions to tackle crime, the 3ie database for systematic reviews and impact evaluations in international development, and ALNAP's Humanitarian Evaluation, Learning and Performance (HELP) Library containing evaluations of humanitarian interventions (Fig. 2).

The next level of the pyramid is evidence mapping which presents the evidence from a database in a structured way with a summary of the main features of that literature. Evidence maps guide users to the evidence and show research commissioners where there are gaps. Research funders around the world should be using evidence maps to inform funding decisions. Maps also increase discoverability. 3ie undertook a map of maps in international development in 2017: 73 maps were found of which 18 were ongoing and a further 42 published in 2015–2016 (Phillips et al., 2017).

Next come evidence platforms. These platforms offer a range of evidence products in a user-friendly way, often with summaries of those studies. Examples are EvidenceAid for humanitarian relief, Eldis for international development in general, the Homelessness Hub, and the Social Care Institute for Excellence.

A key break in the pyramid comes at the next stage. Databases, maps and platforms link users back to the original research papers or summaries of that research. The top three levels—evidence portals, guidance and checklists—enable evidence-informed decision-making without requiring the decisionmaker to look at the research paper. The three levels differ in the agency afforded the decision-maker: evidence portals present the evidence leaving it to the decision-maker to decide, guidance provides recommendation based on the evidence, and checklists present a 'do this' list. These are decisionmaking tools, they do not remove the role of deliberation as discussed by Munro and colleagues for the case of using research for child safety (Munro et al., 2016).

Evidence portals are presented by the various What Works Centres in the UK and USA. The leading examples are IES' What Works Clearing Houses (WWC) and the Education Endowment Foundation's Teaching and Learning Toolkit. These two are best-practice examples of easy to access and understand findings from evidence synthesis on the effectiveness of different teaching, or school and classroom management, approaches. Other examples are the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) 'Best Practice Portal' and the EU-funded Safety Cube of evidence on road safety.

The use of guidelines is best established in health. Internationally the World Health Organization (WHO) produces guidelines which are the basis for the national guidelines adopted in many countries around the world. WHO guidelines are required to be based on high-quality systematic reviews, thus institutionalizing the use of evidence from rigorous synthesis. In the United Kingdom, the National Institute of Clinical Excellence and Social Welfare (NICE) uses systematic reviews both for guidance and to make decisions on eligible expenditures for public spending in the National Health Service. Various UK What Works Centres have started to produce guidelines, such as those on the use of Teaching Assistants from the Education Endowment Foundation (Sharples et al.) and the Neighbourhood Policing Guidelines from What Works for Crime Reduction (College of Policing, 2018).

The case of checklist has been made eloquently by Atul Gawande in *The Checklist Manifesto* (Gwande, 2011). Gwande documents how the use of checklist has reduced 'errors of ineptitude' (failure to use what we know) in everything from flying planes to building skyscrapers. Can such an approach work in other sectors? The experience of the leading knowledge brokers in the What Works movement suggests it can.

The Teaching and Learning Toolkit presents evidence on 34 different interventions such as one to one tuition. The toolkit landing page lists the 34 interventions with three simple metrics: cost (shown on a scale of one to five £ signs), strength of evidence (shown on a scale of one to five lock symbols), and impact. Impact is shown as the months' additional progress a child makes if exposed to that intervention. It is +5 for one to one tuition, meaning that providing a course of one to one tuition has, on average, delivered additional progress equivalent to five months of learning. The best buy is giving the child feedback on their work. It costs very little and is equivalent to an additional 8 months progress. The 'worst buy' is repeating a year which costs a lot, even though the child tends to make less progress than if there had been no intervention at all.

The evidence presented in the toolkit is based on 34 systematic reviews commissioned by EEF. A study by NAO in 2015 found that 64%—that is nearly two-thirds—of schools were using the toolkit to inform decisions about school resources and classroom practice (NAO, 2015, p. 9). That is two-thirds of school in England are using evidence from systematic reviews to inform their decision-making. Such is the power of effective knowledge brokering.

At Campbell we are keen to work with and foster the What Works movement. We would like to see the movement across the world base its evidence standards on systematic reviews like EEF does. This is not universally the case, as shown in the review of evidence standards by David Gough and myself (Gough and White, 2018). And we would like to see the centres commission high quality reviews–of course preferably registered with Campbell or Cochrane as appropriate, which ensures that potential biases are reduced. To this end, Campbell has been working with the UK Centre for Homelessness Impact. We have produced two evidence maps, provided preliminary content for their evidence portal, and they have commissioned three reviews which are registered with Campbell.

There is a need for some coordination here. Reviews review the global evidence and portals are built on that global evidence. It does not make sense for every country to do this work separately and independently. Evidence for Leaning in Australia publishes a Teaching and Learning Toolkit which is simply a reproduction of the EEF toolkit in a nice shade of blue. This is as it should be. Rather than reinvent the wheel, the Australians licence the right to use EEF's work–thus providing income for EEF to maintain and expand the toolkit.

In international development there are many global initiatives which should be taking the lead in building the evidence architecture for their respective sector: nutrition, child violence, financial inclusion, whatever–we need to make the evidence available in all these sectors. Less than 1% of the funds spent by these global funds would be sufficient to build the evidence architecture: great value for money as it reduces the share of the other 99% spent on programmes with weak or no effects, which is likely to be 80% or more of them.

These global initiatives already spend money on research and knowledge brokering, but not in a strategic way to build the evidence architecture. The funds they spend on such activities should be repurposed in a strategic direction. These efforts should be coordinated. Registering reviews with Campbell and Cochrane is one way to achieve this coordination.

One we thing we have learned from evaluations in many sectors is that creating supply is rarely sufficient by itself—attention is also need to the demand side. Say's Law that supply creates its own demand likely does not seem to apply in many cases, and promoting the use of evidence is no exception. So, as suggested in Fig. 2, if we are going to build up the supply side of the evidence architecture then we need also to pay attention to the demand side. This is particularly importance as academic incentives support supply of research evidence, but do not generally reward efforts to have that evidence used in policy. The next section proposes demand side steps in building the evidence architecture. These steps are focused on institutionalizing the use of evidence. I do not discuss other important issues such as stakeholder engagement in setting questions and co-production.

Steps in building the evidence architecture

A first step in building the evidence architecture in a particular sector is to undertake an Evidence Ecosystem Assessment (EEA). This is an assessment of the state of the evidence architecture. It maps which agencies are involved in producing what types of evidence, who is brokering that evidence, and who is using it and for what. The UK Alliance for Useful Evidence has produced an overview of the Evidence Ecosystem which shows the main actors.[7] The ecosystem assessment should engage those responsible for the existing architecture, working to the principle of building on what already exists rather than creating new, parallel structures.

Having identified what is out there the next step is to review or update existing evidence and gap maps (EGMs), or construct a new map or map if suitable ones do not exist. This is a first step in building the architecture and will give a basis for engaging the broader community of users as well as producers.

As described above, the maps increase the discoverability of evidence resources.

The community of users should now be engaged through use of evidence workshops. These workshops review the different types of evidence and their uses. Running the workshops is a useful step for the next stage of undertaking an Evidence Needs Assessment (ENA). This idea is based on an exercise conducted by the UK Cabinet Office—called Areas of Research Interest[8] — in which government departments were asked what research questions they needed answered to inform their decision making. The US 'Foundations for Evidence-Based Policymaking Act of 2017' requires US government departments and agencies to develop a plan which includes 'a list of policy-relevant questions for developing evidence to support policymaking'.[9] The exercise can have systemic effects, making decision makers aware of the fact that it is a good idea to use evidence in their decision-making.

The combination of the ENA and EGMs then identify the priorities in building the lower levels of the demand-side of the evidence period. What primary studies, reviews, and maps are needed? This is where international coordination should come in to avoid duplication in producing reviews and maps.

Once the foundations of the evidence supply pyramid are sufficiently strong then it is time to construct portals, and develop guidelines and checklists. These products will likely be adapted to local contexts and so provide a role for national knowledge brokering agencies.

Once the higher levels of the evidence architecture are in place then commitments can be made to evidence-based budgeting (EBB). Evidence-based budgeting has become common in the United States. It means that money can only be allocated to programmes which are deemed to be evidence-based. The international development NGO has committed to all of its programmes being evidence-based by 2020. Whilst this approach raises issue about the standards to be used in assessing which programmes work, and may fall foul of differences in context or differences in implementation fidelity, it is still a better approach than continuing to fund programmes which, according to our Bayesian principle, most likely don't work.

The final step is to support incentives by instituting the use of evidence assessments of and awards. Results for America publishes an annual assessment of use of evidence called the Invest in What Works Index.[10] The assessment is based on a set of explicit criteria along with a transparent scoring structure. These are developed through a consultative process to ensure buy in, and also to raise awareness as to what agencies can do to increase their use of evidence. Similarly, in the UK, the Institute of Government, the Alliance for Useful Evidence and Sense about Science, have conducted a Government Transparency Check which assessed how transparent government departments are about the evidence behind their policies (Sense About Science, 2018). The assessment is made using an Evidence Transparency Framework (Rutter and Gold, 2015) developed through a consultative process.

In sectors where a systematic process to institutionalize the use of evidence is just beginning it would be premature to undertake an assessment of all agencies, or to publish such an assessment which could create ill will. Hence, in the first years an award will be made for good practice. In later years, as use of evidence becomes more widespread, the assessment of all agencies will be published. This approach is modelled on that of the Mexican national evaluation agency, Coneval. Coneval makes an annual assessment of the quality of the M&E system of government agencies. In the early years it did not publish its assessments but restricted itself to annual awards for good practice in M&E using various categories such as 'generation of evaluations to improve public policy'.[11]

The role of AI, machine learning and Big Data: a fifth wave?

New technologies offer great potential for expanding the production and use of rigorous evidence. Big data provide opportunities for data collection for impact measurement, such as combining satellite data, and rainfall data in assessing agricultural interventions, or data from wearable fitness devices to assess the impact of health interventions or to measure the work effort of rural labourers.

There are also opportunities to improve the production of systematic reviews. Programmes, such as Rayyan and EPPI Reviewer, offer machine learning to assist with screening articles for relevance for inclusion in a review. Cochrane Crowd and Aidgrade use web-based crowdsourcing to screen and code papers with automated meta-analysis in the latter case. The technology is already available for automated living reviews, as algorithms crawl databases for relevant studies, updating maps and reviews as they find them. The human element can come in when discretion or expert judgement is need, such as in guideline production. But having human beings scan articles for relevant text for inclusion is likely a very inefficient way to produce reviews. Adopting these technologies will improve the speed and accuracy of evidence synthesis.

There are also risks. Machines are only as smart as the people they learn from. And the analysis of Big Data needs to be informed by a technical understanding of causal relationships. Correlation is not causation not matter how big the data Elliot et al. (2015). But these are manageable risks which are outweighed by the benefits.

Final word: evidence is the best buy in development

Most interventions don't work, most interventions aren't evaluated and most evaluations are not used. As a result billions of dollars of money from governments and individual donations is wasted on ineffective programmes. Funding research on what works is the best investment we can make. Join the evidence revolution today.

Received: 29 January 2019 Accepted: 15 April 2019
Published online: 07 May 2019

Notes

[1] Campbell's vision for the experimenting society is laid out in D. Campbell (1969). Campbell's full contribution to a range of disciplines can be read in Boruch (2019).
[2] The triple A criteria were proposed in my review of development agency performance measurement presented in White (2005a).
[3] Personal communication from Bill Savedoff.
[4] Results from Google Scholar search: 'systematic review' AND social IN Title. Results screened until five consecutive pages with no eligible studies. Search performed 12 September 2018.
[5] Numbers from 3ie database.
[6] Search on 'systematic review' in Title on ERIC, 28/1/19.
[7] http://www.alliance4usefulevidence.org/assets/Alliance_info_graphic5.pdf
[8] https://www.gov.uk/government/collections/areas-of-research-interest
[9] https://www.congress.gov/bill/115th-congress/house-bill/4174.
[10] https://2017.results4america.org/
[11] https://www.coneval.org.mx/Evaluacion/BPME/GF/Paginas/Buenas-Practicas-2018.aspx

References

Boruch R (2019) Campbell D. In: Delamont S, Atkinson P, Cernat A (eds) SAGE research methods foundations. Sage, London

Cairney P (2016) The politics of evidence-based policy making. Palgrave MacMillan, Springer, London

Cames M et al. (2016) How additional is the clean development mechanism? Analysis of the application of current tools and proposed alternatives. Öko-Institut e.V, Berlin

Campbell D (1969) Reforms as experiments. Am Psychol 24(4):409–429

Campbell D (1988) The experimenting society. In: Overman ES (ed.) Methodology and epistemology for the social science. University of Chicago Press, Chicago

Carvalho S, White H (1994) Indicators for poverty reduction. World Bank Discussion Paper 254. World Bank, Washington D.C.

Center for Democracy and Governance, USAID (1999) Handbook of democracy and governance program indicators Ref: PN-ACC-390. USAID, Washington D.C.

College of Policing (2018) Neighbourhood policing guidelines. College of Policing, Coventry

Connolly P, Keenan C, Urbanska K (2018) The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. Educ Res 60(3):276–291

Elliot J et al. (2015) Making sense of health data. Nature 527:31–32

Evans MC, Cvitanoivc C (2018) An introduction to achieving policy impact for early career researchers. Palgrave Commun 4:Article number: 88

Gawande A (2011) The checklist manifesto: how to get things right. Profile Books, London

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. GAO, Washington D.C.

Gough, D and White, H (2018) Evidence standards and evidence claims in web based research portals. London, Centre for Homelessness Impact

Langer L, Tripney J, Gough D (2016) The Science of using science: researching the use of research evidence in decision-making. EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London, London

Levine R, Savedoff W (2006) When will we ever learn: improving lives through impact evaluation. Centre for Global Development, Washington D.C.

Miguel E, Kremer M (2004) Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities Econometrica 72:159–217

Munro E, Cartwright N, Hardie J and Montuschi E (2016) Improving Child Safety: deliberation, judgement and empirical research. Centre for Humanities Engaging Science and Society (CHESS), Durham University, Durham

NAO (2002) Department for international development performance management —helping to reduce world poverty. The Stationery Office, London

NAO (2015) Funding for disadvantaged pupils. National Audit Office, London

Oakley A (1998) Experimentation and social interventions: a forgotten but important history. BMJ 317(7167):1239–1242

Oliver K, Cairney P (2019) The dos and don'ts of influencing policy: a systematic review of advice to academics. Palgrave Commun 5:21

Oliver K, Pearce W (2017) Three lessons from evidence-based medicine and policy: increase transparency, balance inputs and understand power. Palgrave Commun 3:43

Parkhurst J (2017) The politics of evidence: from evidence-based policy to the good governance of evidence. Routledge, London

Petrosino A, Turpin-Petrosino C, Hollis-Peel M, Lavenberg JG (2013) Scared straight and other juvenile awareness programs for preventing juvenile delinquency: a systematic review. Campbell Syst Rev 2013:5

Pfeffer J, Sutton R (2006) Hard facts, dangerous half truths and total nonsense: profiting from evidence-based management. Harvest University Press, Cambridge

Phillips D, Coffey C, Tsoli S, Stevenson J, Waddington H, Eyers J, White H, Snilstveit B (2017) A map of evidence maps relating to sustainable development in low and middle-income countries evidence gap map report. CEDIL Pre-Inception Paper, London

Rutter J, Gold J (2015) Show your workings: Assessing how government uses evidence to make policy. Institute of Government, London

Scher L, Maynard R, Stagner M (2006) Interventions intended to reduce pregnancyrelated outcomes among adolescents. Campbell Syst Rev 2006:12

Sense About Science (2018) Transparency of evidence: a spot check of government policy proposals July 2016 to July 2017. Sense About Science, London

Sharples J, Webster R and Blatchford P Making best use of teaching assistants: guidance report. Education Endowment Foundation: London

Taylor-Robinson DC, Maayan N, Soares-Weiser K, Donegan S, Garner P (2015) Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance. Cochrane Database Syst Rev 2015:Issue 7

Thyer B (2015) A Bibliography of randomized controlled experiments in social work (1949–2013). Res Soc Work Pract 25(7):753–793

Todd B and the 80,000h team (2017) Is it fair to say that most social programmes don't work? https://80000h.org/articles/effective-social-program/. Accessed 4 Nov 2019

Welch VA 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 metaanalysis. Campbell Syst Rev 2016:7

White H (2002) A drop in the ocean? The International Development Targets as a basis for performance measurement. Appendix 2 in NAO

White H (2005a) Challenges in evaluating development effectiveness. In: Pitman G, Feinstein O (eds) Evaluating development effectiveness. Transaction, London

White H (2005b) The road to nowhere: results-based management in international cooperation. In: Cummings S ed. Why did the chicken cost the road? And other stories on development evaluation. KIT, Amsterdam

Acknowledgements

Thanks are due to Amirah El-Haddad, Danielle Mason, Vivian Welch, and Emmy de Buck.

Additional information

Competing interests: The author declares no competing interests.

Reprints and permission information is available online at http://www.nature.com/reprints

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.