THE

Evaluator

AUTUMN 2019

EVALUATION: A DIVERSE FIELD

UK Evaluation Society

evaluation.org.uk

Contents

Would you like to contribute to the new look Evaluator? Check out notes for contributors on our new look website at: ***evaluation.org.uk***

Editorial Bev Bishop

This year's conference celebrated diversity in evaluation – diversity in terms of the challenges we as evaluators face; diversity of practice and experience across the sectors; diversity in terms of methodology and technique and the synergies and innovations inspired by diverse new configurations of stakeholders working together.

As our President Julian Barr notes, all arms of the evaluation community, from third sector to government were well-represented at Conference, and their main complaint was that they didn't have enough time to talk to each other!

The conference opened with the UK Evaluation Society. Andrew Berry reports back on this, noting the year's key achievements in terms of improvements to our members offer, new training package; and increased visibility, and the Treasurer's and Secretary's support.

Peter Craig's keynote presentation discusses the revision of the UK Medical Research Council's seminal guidance on the development and evaluation of complex interventions. Although targeted towards the healthcare sector, the principles behind the guidance have been widely cited, and influential across the piece.

Three pieces here discuss examples of evaluation in different sectors. Firstly, staying with the health sector, Matt Sharp, Samantha Hinks and Michael Lawrie write about the evaluation of NHS England's New Care Models (NCM) programme and the methodological challenges of evaluation such a complex intervention. In the areas of education, Itora Emembolu, Prof. Rebecca Strachan and Dr Carol Davenport from Northumbria University describe their workshop on an evaluation of an intervention to improve young children's understanding of career options. And in the field of international development Rebecca Murray & Valeria Izzi, address the tricky question of what counts as impact when assessing how research affects development in a complex environment.

The question of diversity in methodology across the piece was explored in a panel session at the conference, summarised here by Michelle Hollier, where participants reflected on the demand for, and challenges of adopting and implementing these approaches in different circumstances and the question of how to assess their appropriateness.

PhD student Robyn Millar, looks at how specific methods, tools and approaches are implemented in diverse programme evaluation settings, focussing particularly on the use of programme theory in the evaluation practice of small third sector organisations (TSOs).

Sarah Brown, Graham Thom and Stefano Conti discuss the benefits and challenges of combining different approaches to evaluating the same programme; in this case the Sutton Homes of Care Vanguard programme aimed at improving the quality of care and resident health and wellbeing in care homes.

A number of contributors discuss specific innovations in methodology.

Adaptive management is a purposely iterative approach to development programmes which draws on 'rigorously embedded' monitoring and evaluation. Elbereth Donovan explains how she combined a contribution analysis (CA) with an explicit process review of monitoring, evaluation & learning (MEL), political economy analysis (PEA) and management activities, and the demands such an approach made on the different stakeholder groups.

Mary Suffield and Ben Collins draw upon their extensive experience of using chatbots, sensors and online communities to argue that digital is a 'need to have' not a 'nice to have' in contemporary evaluation. In a similar vein David Drabble extols the benefits of data visualisation in disseminating evaluation findings.

James Copestake reflects upon designing and testing a qualitative impact protocol (the QuIP) for drawing 'robust generalisations' about the impact of specific rural livelihood strengthening projects in Ethiopia and Malawi.

Matthew McConnachie and Clarissa Samson (LTS International) used a realist approach to evaluate Market-Approaches to Climate Resilience in Ethiopia in both the design and implementation phases and consider the key lessons learned.

A third element of diversity that was a strong theme throughout the conference was diversity among different stakeholders and stakeholder groups.

Rick Davies not only discusses why diversity is important but reports on his development of web app called **ParEvo**.: a web-assisted participatory scenario planning process.

Georgie Parry-Crook re-evaluates the roles of different stakeholders in collaborative evaluation drawing on her experience of co-production and develops a useful typology of different sorts of co-production.

Alison Pollard, Lorraine Wapling and Mark Carew share lessons learned from DFID's Girls' Education Challenge Fund, a scoping study about disability inclusive evaluation processes and systems, while Catherine Kelly reflects on effects of evaluation/ stakeholder relationships on key decision-making in the planning process of a for the evaluation of a initiative to widen participation in higher education in the North East of England.

Next edition

The next edition of the Evaluator will be focussed on ethics. As always, we welcome contributions of think pieces, cases studies, reflections, and book reviews. Please submit your ideas to ***hello@evaluation.org.uk***

From the President
Julian Barr

Arriving at a theme for the Society's annual conference is always difficult. The theme needs to cater to people's interests and bring some coherence, yet not be so focused that it only offers a narrow appeal. Conferences often organise around sectors or methodological schools. Yet evaluation cuts across a broad swath of societal activity and the UK Evaluation Society promotes methodological pluralism. In pondering this, we recognised that a particular strength of the field of evaluation is its very breadth and diversity. So, we set out to celebrate and showcase this in this year's conference.

We wanted to promote the cross-fertilisation of ideas and experiences. We wanted to demonstrate the thematic and methodological diversity that exists in evaluation in the UK. And we wanted to give an opportunity for a wide range of different types of organisations and people working in different roles in the evaluation ecosystem to talk about their work and their evaluation challenges. I think we definitely succeeded in these aims. In both the conference and this issue of The Evaluator we managed to present a solidly representative sample of the exciting evaluations conducted in and by UK organisations.

Two other conference aims that I am particularly pleased we achieved are the strong involvement of public and third sector organisations – including the Charities Evaluation Working group (ChEW), and the use of some different styles of session format. We had a really strong presence from government and departments, with co-ordination from the Cross-Government Evaluation Group (CGEG). CGEG also ran a very well-received panel session on the new, and imminently to be published, Magenta Book guidance on evaluation in government. In terms of formats, both the Magenta Book panel session and the ChEW world café worked extremely well.

In practical terms, the conference was attended by 214 delegates, compared with 177 delegates in 2018, and we continue to be up on our long term average.

UK Evaluation Society

JULIAN BARR – UK EVALUATION SOCIETY PRESIDENT

Delegates were diligent in completing evaluation forms about the conference, and we received very useful and constructive feedback about what worked well, and areas on which we might need to focus attention in 2020. Positive comments included:

"A friendly conference. Very informative and has generated a few ideas to take back to work."

"Variety of presentation styles compared to previous conferences. Quality of presentations / sessions."

"Good exposure to new knowledge; good insight into work across all sectors; incredibly interesting and engaging; well pitched to experienced and new evaluators."

However, several delegates found that there was "not enough structured networking." An area to work on for next year.

My personal thanks to go to the conference organising team, especially Rebecca Adler, who drew up the programme, and to Professional Briefings for administrative support.

We are now planning for next year and I am pleased to confirm the 2020 UK Evaluation Society Conference will be held between Wednesday 3 and Thursday 4 June 2020 at the same venue as the past two years: The Grand Connaught Rooms in central London. We will be very shortly issuing a Call for Abstracts. We're looking forward to a good selection of abstracts and a very well-attended conference in 2020.

UK Evaluation Society AGM 2019

ANDREW BERRY

The 2019 UK Evaluation Society Annual General Meeting (AGM) was opened by President of the Society Julian Barr with a progress report for 2018-19.

Supporting members with CPD and training, networking, information and guidance

Four successful training events were held during the year, attracting international guest speakers on contemporary topics. A second round of Voluntary Evaluator Peer Review (VEPR) was completed and we were delighted to launch the revised second edition of the Good Practice Guidelines for Evaluation. The 2018 conference on The Quality of Evidence from Evaluation was one of our best attended in the past decade and it received very positive feedback from delegates.

Improved functioning of the Society to better support members

As well as the continued efforts of Council and thematic working groups to deliver different aspects of our 2017-19 business plan, we undertook a detailed review of our Society's administration and management arrangements in 2018 and we have introduced a new business model which is being piloted during 2019-20.

Better visibility, growing the Society, leading to better services to members

The new brand and website refresh was conceived to give the Society the best possible platform for improved visibility and growth, and the feedback from members has been overwhelmingly positive.

At the AGM, Andrew Berry, lead for the communications and marketing working group, provided more detail on the development of the Society's new logo, including feedback from the 2018 member survey. The new brand has been rolled out in the form of new designs for The Evaluator, 2019 conference materials, regional and national networks publicity, plus our Guidelines for Good Practice in Evaluation and Framework of Evaluation Capabilities. The new website has been designed with end-users in mind in terms of its look, functionality and content. We are planning for more features to be added, including our new member log-in portal.

Treasurer's Report

Tracey Wond presented a statement to delegates on income for the 2018 financial year of £78,929 and expenditure of £77,367, producing a surplus of £1,546. We will be running a budgeting exercise in 2019-20, accounting for new ways of working within the Society and helping to diversify our revenue streams.

Secretary's Report

Andrew Berry presented a statement to delegates on the number of individual members (111) and institutional members (20) at the time of the 2019 AGM.

During 2018-19, three officers stood down from Vice President, Treasurer and Secretary roles which were subsequently refilled, plus two Council members stood down and eight Council members were (re)elected. The Council complement for 2019-20 is made up of five officers, 14 Council members and four co-opted members.

Conference Awards 2019

Winner of best paper at conference, sponsored by Sage Publishing: Clarissa Poulsen, IPE Triple Line – Evaluating Grand Challenges and Innovation: Approaches, Lessons and Reflections

Winner of best poster at conference, sponsored by Sage Publishing: Graeme Smith and Karen Kerr, Skills Development Scotland – How Skills Development Scotland evaluates the careers service to support young people in making career decisions

Young Evaluator of the Year, sponsored by Ipsos MORI: Niki Wood, Government Partnerships International – presentation on contribution courts

Invitation to join UK Evaluation Society Council working groups

Over the past couple of years we have been evolving a set of functional sub-groups to the UK Evaluation Society Council. These focus on particular areas of Society development and management, and have proved a successful way to work.

Over the past couple of years, we have been evolving a set of functional working groups reporting to the UK Evaluation Society Council. These focus on particular areas of Society development and management, and have proved a successful way to work.

We would like to invite all members to consider joining these groups, to support the running of your Society, and to provide a voice from Society members. The working groups and the aims of each are outlined below.

Group	Aim
Membership	To improve the satisfaction of members the services of the UK Evaluation Society, and grow the number of members of the Society.
National & Regional Networks	To support the development of sustainable national and regional networks that increase the Society's presence at a more local level; helping the Society to be inclusive and responsive to the needs of its members
Communications & Marketing	To keep members informed and engaged with the Society, to help make connections between members, and to promote evaluation-related news and activities more generally
Training & Development	To develop and provide a programme of activities to share expertise and knowledge in evaluation methodologies and approaches to help build capability and capacity in the evaluation community.
Voluntary Evaluator Peer Review (VEPR)	To design and test VEPR and support its adoption as a service for the UK Evaluation Society membership
Evaluation Capabilities	To review, develop and publicise the Framework of Evaluation Capabilities, including launching an online self-assessment tool and working on a revised framework (first published 2012)
Ethics & Good Practice	To design and maintain the Guidelines for Good Practice in Evaluation underpinned by ethical principles; and to promote the Guidelines (first pub-lished 2003, revised 2018) at every relevant opportunity
Strategic Partnerships & Collaborations	To develop partnerships with other organisations to further the aims of the UK Evaluation Society; to enhance our offer to members; to enhance the Society's reputation; and to support long-term sustainability of the Society
Conference	To support the planning and preparation of the annual conference each year; selecting conference theme(s); reviewing abstracts; supporting sponsorship activities; scheduling the days; and liaising with our conference administrators
Young & Emerging Evaluators (YEE)	To engage with and support the professional development of those coming into evaluation, either as students, graduates and post-graduates, or through new or changed employment roles; representing the interests of this group

KEYNOTE PRESENTATION
SESSION SPONSORED BY ICF

From complicated interventions to interruptions in complex systems?

PETER CRAIG, SENIOR RESEARCH FELLOW, MRC/CSO SOCIAL AND PUBLIC HEALTH SCIENCES UNIT, UNIVERSITY OF GLASGOW

The UK Medical Research Council's guidance on the development and evaluation of complex interventions has been highly influential in health services and public health research.

First published in 2000,[1] and revised and updated in 2006,[2] it provides researchers with a straightforward, phased approach to evaluating interventions with multiple, interacting components. The guidance (both 2000 and 2006 editions) continues to be widely cited. It has stimulated the production of more detailed guidance on aspects of the overall process, such as intervention development[3] and process evaluation,[4] and is frequently incorporated into funders' guidance for applicants.

So why update? Critics of the guidance have questioned its underpinning definition of a complex intervention and its focus on researcher-led interventions that are amenable to evaluation via a planned experiment, such as a randomised controlled trial. A number of issues have been identified where the guidance either has little to say, or fails to incorporate recent thinking, such as intervention development, engaging stakeholders, economic appraisal and evaluation, how interventions interact with the contexts in which they are implemented, and systems approaches to evaluation. The decision by the MRC and the UK National Institute of Health Research to commission a revision and update of the guidance is therefore timely.

A group of researchers at the MRC/CSO Social and Public Health Sciences Unit is undertaking the revision. We are taking a systematic and collaborative approach, involving wide-ranging consultation with the research community, via conference workshops, an expert consensus meeting and an online consultation on an early draft. We are also drawing on counsel from an independently chaired scientific advisory group with representation from the funders and researchers from a range of health and social science disciplines. The new guidance will be published early in 2020, but the structure and content have already been settled. In the remainder of this piece I provide an overview of what the guidance will cover, and some of the thinking behind the changes we are making.

1. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. BMJ (Clinical research Ed) 2000; 321:694-6. https://doi.org/10.1136/bmj.321.7262.694

2. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. British Medical Journal 2008; 337:a1655. https://doi.org/10.1136/bmj.a1655

3. O'Cathain A, Croot L, Duncan E, Rousseau N, Sworn K, Turner KM, et al. Guidance on how to develop complex interventions to improve health and healthcare. BMJ Open 2019; 9:e029954. https://doi.org/10.1136/bmjopen-2019-029954

4. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. British Medical Journal 2015; 350:h1258. https://doi.org/10.1136/bmj.h1258

What is a complex intervention?

The first version of the guidance defined complex interventions as those that are built up from a number of components that act both independently and interdependently. The revised version adopted a similar perspective, but remarked that while few interventions are truly simple, there is a wide range of complexity, and drew attention to some of the dimensions along which complex interventions differ from (relatively) simple ones, such as interactions between intervention components, the number of groups or organisational levels targeted by the intervention, or the degree of flexibility or tailoring permitted as the intervention is delivered. Critics of this view argue that while these things make interventions more or less complicated, complexity arises from their interactions with the wider system in which they are implemented. On this alternative view, evaluating an intervention is not just a matter of measuring its impact, but of understanding how it changes the functioning of this system.

Box 1: A simple intervention that is actually very complex

The lucky iron fish is a small fish-shaped iron ingot, placed in a pan of boiling water used for cooking food. Enough iron dissolves in the water to prevent anaemia caused by dietary iron deficiency. Randomised trials in Cambodian villages where anaemia is endemic show that it is effective[5]. Compared with pharmaceutical iron supplements, the lucky iron fish is cheap and can be manufactured locally so that supplies can be maintained.

What could be simpler? Actually, from a systems perspective, there are many sources of complexity. The ingots are fish-shaped so that recipients value them – fish are seen as auspicious in Cambodia and other parts of South East Asia – but not too much. Villagers had to be encouraged to store them in the kitchen and use them for cooking. Effectiveness varies seasonally as the chemical composition of water supplies can change in the wet season reducing the dietary availability of the iron. The fish is less effective in places where anaemia has a genetic rather than a dietary cause.

[5] https://luckyironfish.com/pages/clinical-research
Taking the economic, cultural, geographical and epidemiological context into account, the lucky iron fish is not so simple after all.

Evaluating interventions as interruptions in complex systems is highly challenging, conceptually, methodologically and practically. It is no accident that 'rhetoric urging complex systems approaches is only rarely operationalised in ways that generate relevant evidence or effective policies.'[6] Rather than advocating such approaches as the only worthwhile way to evaluate a complex intervention, the new guidance situates a systems perspective alongside other perspectives. We argue that the choice of perspective should be governed by the questions the research seeks to address, which should in turn reflect the key uncertainties about (for example) the efficacy, effectiveness, implementability or sustainability of the intervention, given existing evidence.

Six core elements

Like its predecessors, the new guidance advocates a phased approach, so that due attention is paid to intervention development, feasibility testing and post-evaluation implementation, as well as to evaluating effectiveness. We have identified six core issues that should be addressed at each of the stages. An intervention that works in one **context** may be less effective or even harmful in others, due to differences between contexts in the resources available to deliver the intervention, in the characteristics of the recipient population, or in the range of co-interventions already in place. Taking account of how variation in context may moderate the effect of an intervention is therefore a key consideration from development right through to large scale implementation.

Context is an important element of a **programme theory** – a theory that describes how an intervention is expected to achieve impact. A good programme theory will describe the components of the intervention and how they interact with one another, the mechanisms through which change is expected to occur, how context moderates the effect of the intervention, and how the intervention may change the context in which it is implemented. A ban on smoking in enclosed public places may have an immediate effect on exposure to second-hand smoke, by preventing smokers from lighting up in public. By reducing the visibility of smoking, a ban may also have a longer term effect on uptake or progression to regular tobacco use.

Stakeholders should be involved throughout the development and evaluation process. We define stakeholders broadly to include intervention recipients as well as those working in any capacity – organising, planning, funding, etc. – to deliver the intervention. Engaging stakeholders thoroughly helps to ensure that the questions being asked by the evaluation matter, and that the evidence is taken up and used beyond the immediate context of the research study.

Development and evaluation of interventions usually builds on a substantial base of existing evidence, such as evaluations of similar interventions in other settings. The focus of any new cycle of research should be on the key uncertainties defined by what is already known, and by what stakeholders identify as important. Judgements about what are the key **uncertainties** inform the framing of research questions, which in turn govern the choice of research perspective.

Some interventions only work if they are delivered in a tightly specified way (for example where a defined dose of a drug is required to clear infection). But many interventions work better if some flexibility or tailoring is allowed to improve their fit with a particular context. **Refinement** of an intervention in the light of experience is good practice, so long as the refinements are consistent with the programme theory, and any changes made are recorded and included in evaluation reports.

Complex interventions will nearly always be costly to deliver, or will impose costs on individuals or organisations. **Economic considerations** should therefore always be addressed. Economic evaluation comparing costs and outcomes of an intervention should adopt a broad perspective (for example a societal perspective, rather than the perspective of a single policy sector such as healthcare), and use a broad framework such as cost-benefit, rather than a narrower cost-effectiveness or cost-utility framework. Economic considerations can also be used to help decide whether and how to proceed with costly evaluation activities by formally comparing the cost of research with the expected value of the evidence in terms of reducing uncertainty in future implementation decisions.

Conclusion

The new guidance will be published in Spring 2020. It will take its place alongside other recently-developed guidance, such as INDEX for intervention development,[7] and the NIHR-CIHR guidance on taking account of context in population health intervention research.[8] Further guidance on pilot and feasibility studies[9] and the adaptation of interventions for implementation in new contexts [10] is under development. It is no surprise that, given its origins, guidance of this kind has been used most widely in the health sciences, but it is equally applicable to complex interventions in other policy sectors. We hope that by taking into account realist and systems perspectives, the new MRC guidance will have cross-disciplinary, cross-sectoral appeal.

6. Rutter H, Savona N, Glonti K, Bibby J, Cummins S, Finegood DT, et al. The need for a complex systems model of evidence for public health. The Lancet 2017; 10.1016/S0140-6736(17)31267-9. https://doi.org/10.1016/S0140-6736(17)31267-9

7. See note 3.

8. Craig P, Di Ruggiero E, Frolich KL, Mykhalovskiy E, White M, on behalf of the Canadian Institutes of Health Research (CIHR)–National Institute for Health Research (NIHR) Context Guidance Authors Group. Taking account of context in population health intervention research: guidance for producers, users and funders of research. Southampton; 2018. https://doi.org/10.3310/CIHR-NIHR-01

9. Moore, L, Hallingberg B. Wight, D, Turley R, Segrott J, Craig, P. et al. Exploratory studies to inform full-scale evaluations of complex public health interventions: the need for guidance. Journal of Epidemiology and Community Health 2018 72(10), pp. 865-866. (doi:10.1136/jech-2017-210414)

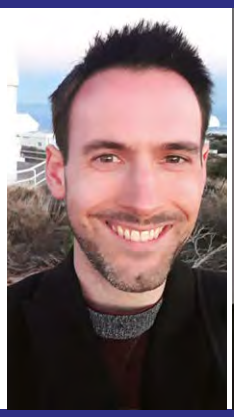
10. Evans, R. E. et al. When and how do 'effective' interventions need to be adapted and/or re-evaluated in new contexts? The need for guidance. Journal of Epidemiology and Community Health 2019 73(6), pp. 481-482. (doi:10.1136/jech-2018-210840)

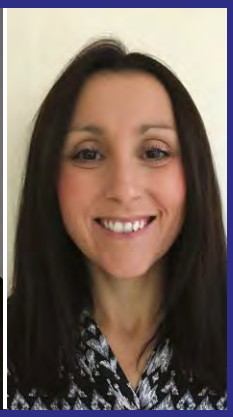
New Care Models: Evaluation of a Complex Programme of Integrated Care

MATT SHARP, ANALYST, NHS ENGLAND AND NHS IMPROVEMENT

MIKE LAWRIE, IPSOS MORI, FORMERLY SENIOR ANALYTICAL LEAD AT NHS ENGLAND AND NHS IMPROVEMENT

SAM HINKS, SENIOR ANALYTICAL MANAGER, NHS ENGLAND AND NHS IMPROVEMENT





EVALUATING COMPLEXITY IN THE HEALTH SECTOR – CONFERENCE PRESENTATION SESSION SPONSORED BY THE HEALTH FOUNDATION

The Health Foundation

Introduction

In 2015, 50 sites across England were selected to be vanguards, as part of NHS England's New Care Models (NCM) programme. These sites each implemented one of 5 different new care models[11]. This article sets out some of the learning from the evaluation of the vanguards that were focused on improving vertical and horizontal integration of GP, hospital, and community services, as well as the vanguards that centred on care homes.

Evaluation Approach

The evaluation of the vanguards was multifaceted, mixed-methods, and framed around a rapid cycle approach. It was supported by substantial funding, which was used to resource both local and national-level evaluations. Vanguards created logic models to set out their anticipated activities and outcomes for their new care models, used local evaluation funding to commission studies (from academic groups, commissioning-support units, and consultancy firms), and reported local metrics which focused on outcomes the vanguards considered particularly relevant to their local programme[12]. Some vanguards also undertook evaluation work in-house. Vanguards were largely free to choose what they wanted to evaluate. The national-level evaluation focused on key metrics relating to hospital activity, studies of key interventions that were implemented across multiple sites, and finally a synthesis of findings from all the local evaluations.

Key Challenges

The key methodological challenges identified from the NCM evaluation echo those of the recently published systematic review by Kadu et al[13], and similar findings are noted in other reviews of integrated care[14,15,16]. These challenges matter because using poor quality or misrepresented evidence risks the implementation of ineffective policies. The key challenges the NCM evaluation faced included:

- Having enough time available – for the interventions of interest to be fully implemented, and to undertake rigorous evaluation of impacts
- Forming a suitable comparator group
- Synthesising a large amount of evidence from multiple sources – unlike most individual evaluations, this was a particular challenge given the large scale of the NCM evaluation programme

11. A More information on each of these models is available here: https://www.england.nhs.uk/new-care-models/about/
12. The University of Manchester undertook an independent review of the locally commissioned evaluations: https://www.research. manchester.ac.uk/portal/en/publications/investigating-locally-commissioned-evaluations-of-the-nhs-vanguard-programme(3c8cfbf6-52de-4639-b715-2df5627c105e).html
13. https://www.ijic.org/articles/10.5334/ijic.4675/
14. https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-018-3161-3
15. http://eprints.whiterose.ac.uk/136498/
16. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4491323/

Each of these is discussed in more detail below.

Sufficient time

Implementation of integrated care interventions can involve substantial restructuring of services, including changes to patient pathways, planning, drawing up new contracts, and hiring of new staff. Once implemented, it may take a while for patients to be routed along the new pathway, and to experience any benefit from it compared to the old pathway. As noted by Kadu et al and Baxter et al[17], this can mean the full effects of an intervention may not be seen until 3 to 5 years after the start of implementation.

This proved a particular challenge in the NCM evaluation due to the evolving nature of the programme, and the appetite for early findings to inform policy. To some extent this was addressed by considering measurements of interim indicators. Where changes in interim measures are known to strongly correlate with changes in the outcome of interest, there is less of a need to wait to see changes in the outcome. This is particularly the case if an intervention is perceived to be low-risk and low-cost, as there is a less of a risk of an adverse outcome and/or wasted resources. However, identifying such interim indicators, and then establishing metrics and data collection, could have been done more successfully in the NCM evaluation had there been stronger up-front theoretical work, drawing on existing literature, as well as earlier and more detailed qualitative and descriptive work to understand what the new interventions were actually doing.

Challenges around suitable comparator groups

Accessing data to form a suitable comparator group proved difficult for numerous interventions across many vanguards, with a key barrier perceived to be laws around information governance (IG), as well as practical issues that reduced data accessibility, such as incompatible computer systems. However, there were sometimes also barriers in terms of some evaluators having not sufficient expertise of different quantitative methods and/or time to access the required data.

Despite these challenges, there were many achievements. For example, the Improvements Analytics Unit (a collaboration between NHSE/I and the Health Foundation) have undertaken matched control studies of multiple vanguards[18]. However, the use of these control groups cannot account for all potential confounders. For example, multiple other programmes were being implemented/run at the same time as the NCM programme, including other programmes relating to integrated care, such as the Pioneers[19]. Other changes in NHS policy and funding may have also acted as confounders, making it difficult to establish the true drivers of change even when positive impacts were identified.

Synthesising a large amount of evidence

Upon completion of the programme, close to 200 evaluation reports were received from the vanguards, covering a diverse set of interventions and outcomes including community mental health, prevention programmes, medicines reviews, and emergency hospital activity. Many reports discussed multiple interventions. In addition, there were 6 quarters worth of around 600 different local metrics, plus evidence from other sources (such as quarterly reviews and presentations) that were used to further triangulate findings. This large and diverse volume of material proved challenging to evaluate and synthesise, requiring the development of a common coding framework to provide structure to the findings. In hindsight, a more focused approach, drawing on theory and published evidence to identify key interventions and key questions, may have been more valuable.

Conclusion

The challenges outlined here are in line with those experienced by other evaluations of complex interventions. A greater emphasis on understanding the theory and existing evidence before undertaking the NCM evaluation would likely have enabled a more focused approach. Similar future evaluations should seek to engage with policy-makers as early as is possible, ideally well before anything is implemented.

Despite these challenges, the local vanguard teams themselves valued the support provided by the national evaluation team[20]. And crucially, the pragmatic approach taken by the NCM evaluation allowed for many useful pieces of learning from the NCM programme, particularly regarding implementation evidence, which has been fed into the NHS's Long-Term Plan.

17. https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-018-3161-3
18. https://www.health.org.uk/funding-and-partnerships/our-partnerships/improvement-analytics-unit
19. http://piru.lshtm.ac.uk/projects/current-projects/integrated-care-pioneers-evaluation.html
20. Satisfaction with the evaluation support provided was higher than any other area; see p.25 of this NAO report: https://www.nao.org.uk/wp-content/uploads/2018/06/Developing-new-care-models-through-NHS-Vanguards.pdf

Designing Effective Evaluations for Applying Scientific Academic Research to Career Based Interventions with Younger Children

ITORO EMEMBOLU, REBECCA STRACHAN, CAROL DAVENPORT



Itora Emembolu, Prof. Rebecca Strachan and Dr Carol Davenport from Northumbria University describe their interactive workshop on the theme of Evaluation Collaborations – learning and working together to enhance practice.

Research highlights that to broaden children's horizons and open up future opportunities, they should have learning experiences on real world applications and careers from an early age, (Chambers et al., 2018). In parallel there is currently a strong emphasis to ensure academic research impacts on wider society. This is evidenced by the 25% assessment allocated to measure impact within the Research Excellence Framework (REF 2021) for quality of research in UK Higher Institutions. This article describes a research project which integrates these two elements to investigate whether academic research on materials science can be successfully deconstructed into an effective career-based intervention with younger children.

Drawing on New Philanthropy Capital's 4 pillar approach for effective impact evaluation (Kazimirski & Pritchard, 2014), this article outlines the design and implementation of an age appropriate evaluation to examine the impact of this intervention on young children's career perceptions and subject knowledge.

1. **Map your theory of change.** Mapping a theory of change is important because it shows pathways that indicate why and how a change should occur. The theory of change developed in the research is part of a wider theory of Change used by NUSTEM to improve young people's uptake of STEM disciplines. NUSTEM is a collaborative STEM outreach group that works with young children and their key influencers.

2. **Prioritise what you measure.** Since the research was concerned with providing career learning experiences for the children, it was necessary to explore the aspirations of the participants, their knowledge of the career promoted and inclination towards that career. This was to help evaluate if the intervention was having any impact on their career aspirations or subject knowledge.

3. **Choose your level of evidence.** The level of evidence collected was a pre- and post-intervention comparison, consistent with the Office of Fair Access (OFFA)'s level 2 standard for evaluating changes in participants perception and understanding (Crawford et al. 2017).

4. **Select your sources and tools.** This aspect of the research involved identifying the different stakeholders to engage, which disciplines of academic research to work with, and which primary schools and cohort classes.

Design and Delivery Approach

The workshop was designed using the cyclical four stages of action research; reflect, plan, act and observe (Leitch & Day, 2000). The action research approach is a valuable approach for an evaluator to adopt because it constantly feeds back in loops. This enables lessons learnt to be incorporated in further iterations during the course of a research.

- **The reflection stage** included identifying and engaging the different stakeholders. It also involved identifying features based on academic(s) research that should be incorporated into the workshop.
- **The planning stage** had to do with co-designing and co-creating the structure and content of the intervention workshop by both academics and outreach specialists. This comprised defining the objective(s) of the workshop; identifying and designing the activities used; designing how the workshops should be evaluated and the protocols to identify, engage and on-board beneficiary schools and cohorts of young people. Repeated stakeholder engagement enabled proper alignment of the objective of the academic based workshop with the activities included in the workshop
- **The acting stage** was mainly concerned with the implementation and delivery of the workshop. This comprised logistics involved in the delivery, engagement with the young people and data collection.
- **The observe stage** was a review of how the workshop was implemented. Aspects of the workshops that worked well and those that did not were identified. This stage also included the evaluation of the workshop and how data collected was analysed.

The cyclical nature of action research means that lessons learnt from the workshop can be incorporated and influence the design and implementation of subsequent workshops.

The study was carried out in North East England with young people aged 7 – 11 years and comprised a pilot stage (n=54) and main study (n=102).

Evaluation Tool

Despite many available evaluation tools, there are limited published tools appropriate for younger children (Padwick et al., 2016; Dele-Ajayi et al., 2018) that could be embedded within the workshops. Some of the considerations taken into account in searching for an instrument were: age appropriateness of tool, short completion time given the average length of a workshop was one hour, the tool needed to be non-intrusive because it had to be accommodated within the workshop delivery, and the tool needed to elicit prior- and post-workshop understanding from the participants. Many of the available tools did not meet these requirements so this led to the need to create an appropriate tool. The study adapted aspects of the Teaching Assessment in Primary Science (TAPS) from Bath Spa University to create the main evaluation tool; the knowledge map (see Figure 1).

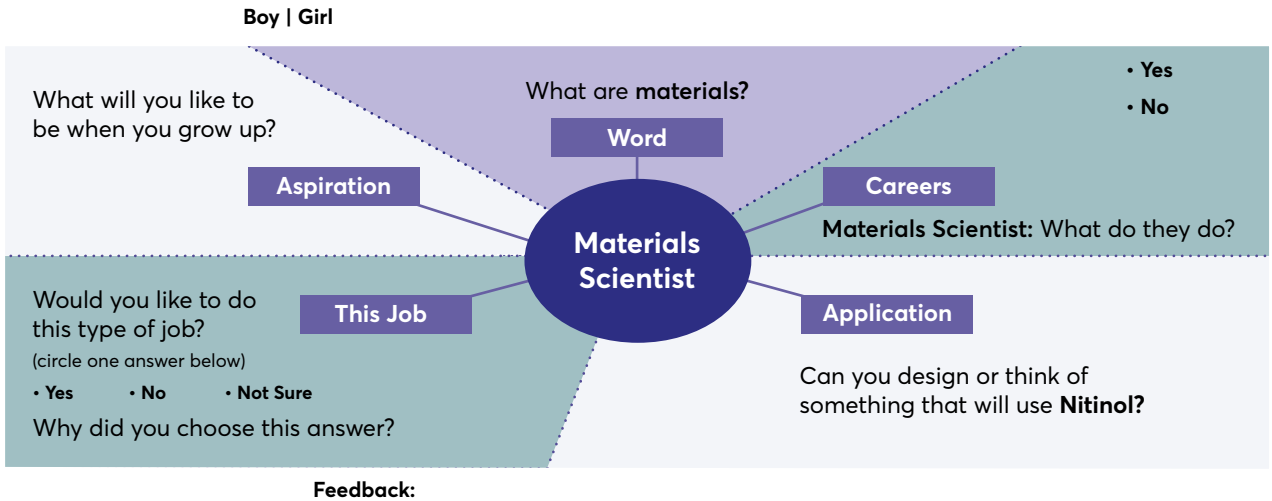
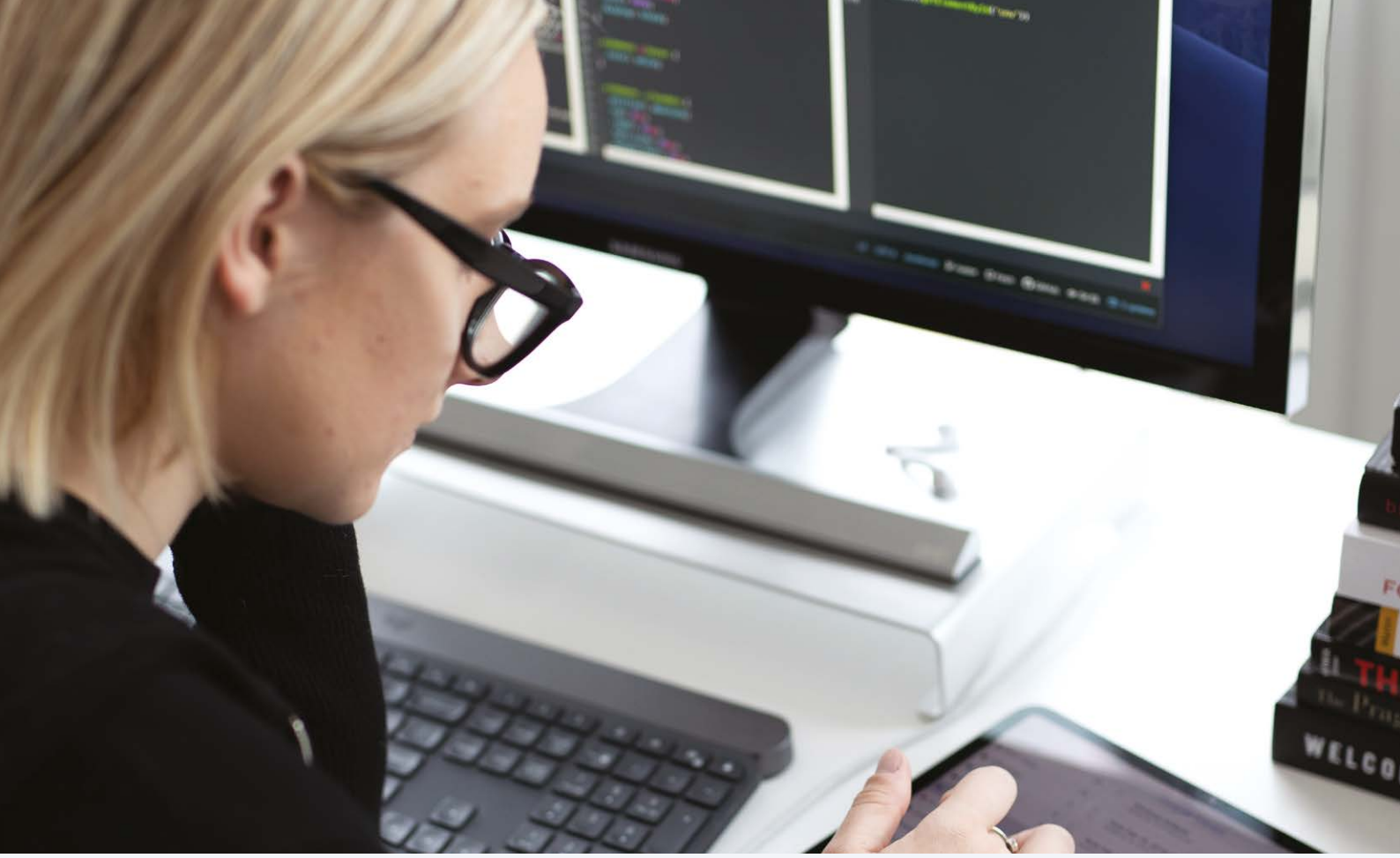


FIG 1. KNOWLEDGE MAP IN A MATERIAL SCIENCE INTERVENTION



Findings

There was an increase in young people's knowledge of the career promoted post-intervention compared to pre-intervention. Whilst the young people indicated they enjoyed the workshop, they were not inclined to aspire to the career. This finding resulted in an adaptation of the tool between the pilot and main study to understand the reasons behind these choices. On further examination, results showed that despite enjoying the workshops, many of the young people already had an idea about the kind of career they wanted to pursue. This has influenced the focus of the team's work with younger children to focus on how to ensure that young people have a good knowledge about a range of careers, so they can keep their options open and therefore, can make informed decisions about their future study and career pathways.

Key Lessons:

- *The iterative nature of action research utilising evaluation tools provides useful insights which can be used to effectively refine interventions and evaluation tools for future deliveries.*
- *The evaluation approach and tools for younger children need to be carefully designed to be age appropriate.*
- *Data from the evaluation provided useful insights into each child's reasoning behind their career preferences or choices.*

- **Chambers, N., Kashefpakdel, E. T., Rehill, J., & Percy, C. (2018).** Drawing the Future: exploring the career aspirations of primary school children from around the world. Education and Employers, London.
- **Crawford, C., Dytham, S., & Naylor, R. (2017).** The Evaluation of the Impact of Outreach: Proposed Standards of Evaluation Practice and Associated Guidance. Bristol: Office for Fair Access.
- **DeWitt, J., & Archer, L. (2015).** Who aspires to a science career? A comparison of survey responses from primary and secondary school students. International Journal of Science Education, 37(13), 2170-2192.
- **Kazimirski, A., & Pritchard, D. (2014).** Building your measurement framework: NPC's four pillar approach. London: New Philanthropy Capital.
- **Leitch, R., & Day, C. (2000).** Action research and reflective practice: Towards a holistic view. Educational action research, 8(1), 179-193.
- **NUSTEM** https://nustem.uk/
- **REF2021** https://www.ref.ac.uk/
- **Riegle-Crumb, C., Moore, C., & Ramos-Wada, A. (2011).** Who wants to have a career in science or math? Exploring adolescents' future aspirations by gender and race/ethnicity. Science Education, 95(3), 458-476
- **Scott, C. M. (2016).** 'To Be a Scientist Sometimes You Have to Break Down Stuff about Animals': Examining the Normative Scientific Practices of a Summer Herpetological Program for Children. International Journal of Science Education, Part B, 6(3), 325-340.

The UK Evaluation Society supports the future of evaluators by promoting and improving the theory, practice, understanding and utilisation of evaluation.

Assessing the impact of research for international development: What does impact look like and how do we measure it?

REBECCA MURRAY, LTSI & VALERIA IZZI,
INTERNATIONAL DEVELOPMENT ADVISER,
UNIVERSITY OF EDINBURGH

An increasing amount of research in the UK is funded through Official Development Assistance (ODA), with the launch of the Global Challenges Research Fund (GCRF) in 2016 having been a game-changer. At a time of science budget cuts and Brexit-related uncertainty, one immediate consequence is that universities across the UK - including those with limited experience of development research so far - have been busily gearing up to access these new sources of funding. For many researchers, GCRF represents the first foray into Research for Development (R4D).

So far, requirements to demonstrate impact for GCRF investments have been more relaxed compared to traditional development funding streams. However, as time goes by, this is likely to change - not least because GCRF will need to fend off concerns that 'research as usual' is being cosmetically rebranded as R4D, and that aid resources are diverted away from their intended anti-poverty focus (Manji & Mandler, 2019; Ritchie, 2019).

Designing a Monitoring, Evaluation and Learning (MEL) system for R4D is not merely a technical matter to be outsourced to MEL specialists – rather, it goes to the core of unresolved questions about the development impact of research: what counts as impact? Impact for whom? And, ultimately, is R4D investment a good use of the UK development budget?

R4D pathways to development impact

Most R4D projects set out to follow a fairly standard pathway to impact - by which research is produced and disseminated to relevant stakeholders, who then put research into use for policy and practice (Morton, 2015). This 'research into use' model has the potential to lead to large-scale impact – but poses significant MEL challenges.

Policy (and practice) change is a non-linear, highly complex process shaped by a multitude of interacting forces and actors, which runs on longer timelines than the average research project. It is difficult to isolate the impact of a particular intervention and to determine the link between research activities and policy change. Methodologies such as experimental and quasi-experimental impact evaluation are unsuitable for policy influencing work, given that it is difficult to establish a plausible counterfactual. In addition, influencing work is most effective when carried out as part of something bigger (alliances, coalitions and networks), which compounds the challenge of judging the specific contribution of any one project (Jones, 2011). Having to 'prove' contribution can also create perverse incentives for the research team, which are encouraged to spend time and resources to chase proof of impact – for example, quotes by policy-makers - rather than focusing on impact itself.

MEL frameworks for R4D are often biased towards quantitative indicators that focus on research production and uptake, looking at the quantity of research outputs as well as their quality in purely academic terms. As we move from research production to research uptake and use, frequently used indicators look at how research is directly quoted in policies or other similar documents. Counting citations is appealing for its relative simplicity, but it presents problems. Research quotes "may be tactical, to justify a political decision that has already been made and over which the actual research, in fact, had no actual influence" (Jones 2011: 6). Moreover, "research will rarely be used directly, but often influences policymakers more gradually and in an amorphous way through 'enlightenment', by providing concepts and ideas" (Jones, 2011: 6). Yet, while it makes sense to see research use more as adaptation and transformation rather than straightforward application, there comes a point "where such refinements are so extensive that it is no longer legitimate to refer to this process as 'research use' at all". (Nutley et al, 2007: 59).

With a much broader and more mature range of R4D investments now underway (or completed), we see that not all projects fit squarely into the 'research into use' model. In some cases, impact comes directly from development activities that are carried out in parallel with the research. Indicators for this kind of pathway are closer to 'traditional' development measures of success. It is somewhat easier for this pathway to establish a counterfactual and attribute these changes to the intervention. The results are very tangible in terms of impact storytelling, and thus more easily available for communications purposes. However, impact for this prototype tends to be small scale and it can be difficult to demonstrate scalability, which is what many funders ultimately want to see.

R4D researchers often talk of other types of impact, of a more intangible nature, resulting from the process of research itself. Indeed, the 'gold standard' of impactful international development research involves equitable north-south partnership, interdisciplinary collaboration and co-production with non-academic actors – to maximise the quality of scientific output, enhance the sustainability of development outcomes and strengthen capacity for further research and knowledge exchange. Emerging evidence seems to point to the fact that such interactions can, in themselves, generate development impact. However, such "impact by process" is probably the most difficult to capture and tends to fall outside the radar of standard MEL frameworks.

MEL implications

The metrics and methods traditionally used to assess the success of academic research differ significantly from those used by mainstream development interventions. Expecting ODA-funded research to simply 'tick the boxes' of both – or spontaneously find an acceptable middle ground – is unrealistic.

These MEL challenges are not insurmountable – but in order to address them, it is important to see MEL in the broader context of impact, rather than reduce it to reporting requirement. Using a Theory of Change as a starting point for developing a MEL framework can play a crucial role in addressing the challenges discussed above. Theory of Change can help R4D research teams to better conceptualise their end goal in terms of securing positive change in the lives of targeted groups of poor or disadvantaged people, avoid an overly strong focus on academic metrics, and more systematically map the context in which they are working and the assumptions underpinning their approach. Such a clear understanding and articulation of intended development impact should be used to shape all aspects of project operations and strategy, and relatedly become the foundation of a meaningful and strong MEL system.

"Theory of Change can help R4D research teams to better conceptualise their end goal in terms of securing positive change in the lives of targeted groups of poor or disadvantaged people."

Impact is repeatedly described by researchers as non-linear, unpredictable, and to a certain extent serendipitous, with many critical factors beyond the control of the project team. A challenge for R4D MEL systems is, therefore, how to be open to capturing unexpected (positive or negative) impact. Crucially, not all unintended impacts are created equal – they have different degrees of 'knowability'. The use of a Theory of Change - contextually relevant, informed by the views of different local stakeholders, and periodically revised - can inform the creation of a MEL system that indicates 'where to look' for possible unintended effects.

About the authors

Rebecca Murray is a Senior Consultant with LTS International. She was the Impact Manager for the Ecosystem Services for Poverty Alleviation (ESPA) programme, and has a background in impact practice in the UK and internationally, having worked in the private, Government, and NGO sectors.

Valeria Izzi is research and evaluation consultant, with a background working in international development with UN agencies, NGOs and academia. She was the Impact and Learning Specialist in the ESPA Directorate.

Boaz, A., Fitzpatrick, S. and Shaw, B. (2008), "Assessing the impact of research on policy: a literature review", Science and Public Policy, 36(4): 255–270.

Ghosh, P. (2019), "Research body to fund humanitarian efforts", BBC News, January 23rd.

Jones, H. (2011), A guide to monitoring and evaluating policy influence, London: Overseas Development Institute.

Manji, A. & Mandler, P. (2019), "Parliamentary Scrutiny of Aid Spending: The Case of the Global Challenges Research Fund", Parliamentary Affairs, 72(2): 331-352.

Morton, S. (2015), "Progressing research impact assessment: A 'contributions' approach", Research Evaluation, 24(4): 405-419.

Nutley, S, I Walter and H Davies (2007), Using Evidence: How Research Can Inform Public Services, Bristol: Policy Press.

Ritchie, E. (2019), ODA for Research & Development: Too Much of a Good Thing?, Centre for Global Development, 11 March: https://www.cgdev.org/blog/oda-research-development-too-much-good-thing

Diversity of Methodology: debate on modern evaluation methods

MICHELLE HOLLIER, DIRECTOR OF RESEARCH
AND EVALUATION AT WINNING MOVES

Michelle describes the panel session delivered by members of the steering group for the UKES Midlands Regional Network on issues and opportunities relating to innovation in evaluation.

We live in a world of constant change, where skill sets can become obsolete in just a few years. Professionals need to consistently upgrade and develop themselves. In this panel session, hosted by Karl King (Director of Service Development at Winning Moves), I, along with two other members of the steering group for the UKES Midlands Regional Network (Tracey Wond, UKES Treasurer and an academic at the University of Derby and Hamayoon Sultan, Independent Evaluation Consultant; formerly Global Monitoring and Evaluation Lead at Islamic Relief Worldwide) reflected on the implications of this for the evaluation community.

In a lively discussion, we considered:

- The demand for new methods and new approaches in evaluation;
- The challenges of implementing these in practice (what new methods there are and how we use them);
- How we can assess whether new methods and approaches are working appropriately, and;
- The conditions needed for innovative approaches to be adopted successfully in an evaluation.

What were the new methods and approaches we were debating? We discussed the use of new ideas and techniques in evaluation, through reflecting on our first applications of new methods such as QCA or alternative Difference-in-Difference designs and our experiences of doing this. We considered the opportunities and challenges in integrating approaches from other disciplines, such as co-production. We also discussed how we respond to encouragement from commissioners and colleagues to innovate, whether this is through specifying particular approaches or simply through explicitly welcoming 'innovation'.

There were some areas of consensus. We agreed that techniques (whether new or not) need to be 'appropriate' to the evaluation question and the context in which the evaluation is being conducted. We also acknowledged that CECAN (and others) were promoting new methods as a way of tackling complexity.

We reflected on the fact that innovation is not just about new methods but rather how we as evaluators do our jobs, the questions we ask and the quality of our analysis. Moreover, that innovation was a function of the relationship between the evaluator and the funder / commissioner. Here, we considered that a good relationship between evaluator and funder / commissioner is important to support the use of new methods and approaches. This relationship needs to be built on trust and openness. There needs to be an acknowledgement of where the evaluation is treading 'new ground' and a willingness to embrace the risks that this might entail, that some things might not proceed to plan or work.

However, key ideas that emerged from the debate were with regards to our role as evaluators:

1. A lot of new approaches are being developed deep in particular specialisms. This sets the bar really high for existing and emerging evaluators – exaggerating any cases of 'imposter syndrome' already present.

2. The use of new techniques often requires the development of new language or ways of understanding what is being done and what is being learnt. This can place a capacity building demand on the evaluator.

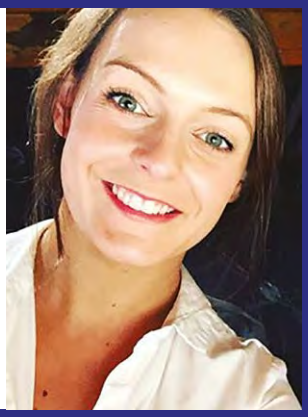
3. The application of techniques takes time in the early stages. It is therefore helpful if the evaluator is well-resourced, but there also needs to be space for thinking and reflection.

Two overarching areas of reflection were therefore on the structural aspects and developmental aspects of evaluation that can support new methods. Structurally, we need to work collaboratively with policy actors and commissioners to embrace evaluation. Developmentally, we posed a question of how we support the CPD of evaluators to keep up with innovations in the way we do evaluation and the methods we use.

"We look forward to continuing the discussion at events for the Midlands Regional Network; please do join us."

Putting (programme) theory into practice in small third sector organisations tackling social issues: exploring current practice

ROBYN MILLAR, DEPARTMENT OF MANAGEMENT SCIENCE, UNIVERSITY OF STRATHCLYDE BUSINESS SCHOOL



Introduction

What is 'research on evaluation practice'?

In practice-oriented fields such as programme evaluation, it becomes really important that we step back and take a closer and more systematic look at that practice in order to understand what is happening, why it is happening, and how we can improve it (Coryn et al., 2016). That not only means looking at how evaluation is conducted but at how the specific methods, tools and approaches that underpin the field of evaluation are being implemented in diverse programme evaluation settings. This research aims to explore the use of programme theory in the evaluation practice of small third sector organisations (TSOs).

What is evaluation practice?

Essentially, evaluation practice is the 'doing' of evaluation. But to conduct research on evaluation practice, we can define evaluation practice in a number of ways which affects the subject of the research as well as the analytical and methodological focus. There are several questions related to the question, 'what is evaluation practice?':

- Who is conducting evaluation? Are they trained in evaluation methodologies or not? If not, what guides their practice?
- What kind of evaluation practice activities are we interested in? Is it systematic studies on some aspect of a programme (Mark, Henry and Julnes, 2000), or the use of evaluation findings, or the evaluative thinking skills required to define, frame, and carry out systematic evaluation activities?
- How does evaluation practice happen? Is practice considered as the application of a methodology/ approach, or as the language, habits, and judgments of those practicing evaluation, i.e. a practitioner focus?
- How do we consider the role of context, i.e. the social, political, and organisational context in which evaluation is taking place?

Evaluation practice in small third sector organisations (TSOs)

In small TSOs, evaluation practice is characterised by several features. First, many of those conducting programme evaluation in small TSOs are not formally trained in programme evaluation, rather they are professionally trained fields such as social work, youth work, or community development. Moreover the small size of many TSOs mean programme evaluation is conducted alongside other roles such as service delivery and is constrained by financial and time resources (Ellis and Gregory, 2009). Second, complex funding and accountability environments of small TSOs present additional demands for conducting programme evaluation. Third, small TSOs often implement programmes to tackle challenging social issues, which are complex in terms of being caused by and manifested in multiple problems (Valentine, 2016) making the evaluation of change in outcomes as a result of a programme challenging. Last, given the preceding challenges, evaluation practice in small TSOs often has to take several forms and serve several purposes simultaneously e.g. programme improvement, accountability, and learning (Ellis and Gregory, 2009; Carman, 2011) in order to ensure the long-term sustainability of these organisations.

Programme theory

This research focuses on the use of programme theory, i.e. the explicit use in evaluation of the assumptions underlying how and why programmes work to change outcomes. Programme theory has increased in interest in the evaluation field over the last 30 years resulting in a plethora of approaches to using programme theory in evaluation, including theory-based evaluation, realist evaluation, theory-of-change evaluation, and logic modelling. Broadly speaking, these approaches aim to understand how (programme process/theory-of action) and why programmes contribute to changes in outcomes (programme impact/theory-of-change) as shown in Figure 1.

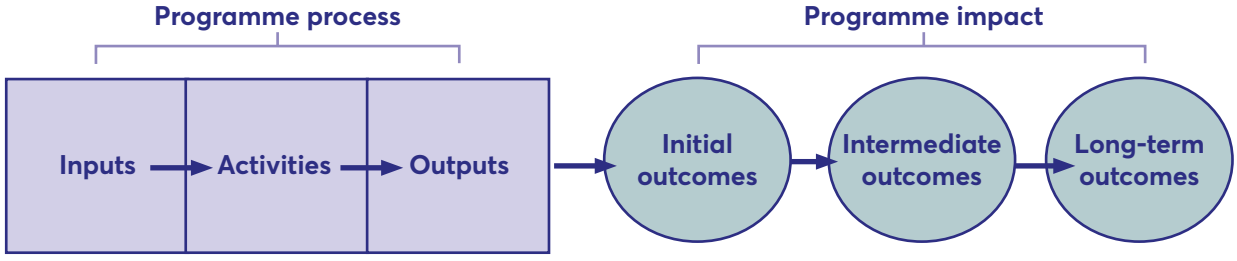


FIGURE 1 PROGRAMME THEORY - ADAPTED FROM DONALDSON (2007)

However, despite the increases in the number of publications on the use of programme theory and related approaches, the perspective on evaluation practice reflected does not necessarily align with how programme evaluation practice is characterised in small TSOs. First, much of the research is from the perspective of trained evaluators (Coryn et al., 2011), rather than those not trained in evaluation. Second, much of the focus in the literature is on the use of programme theory in systematic evaluation activities. However, this perspective does not account for how programme theory can be used to stimulate evaluative thinking for example, which is an important skill to develop for those not formally trained in evaluation. Third, research in this area tends to focus on reported methodological implementation, ignoring the practitioner perspective in the use of programme theory (Coryn et al., 2011; Marchal et al., 2012). This is an important perspective given that practice is not simply theory enacted (Schwandt, 2003). Last, the use of programme theory often focuses on its ability to represent and examine the evaluand. This means that the context of evaluation practice is often omitted in favour of learning about a specific programme.

Research objective and approach

This study explores current practice in the use of programme theory in small TSOs from a practitioner perspective. 23 semi-structured interviews were carried out with participants from small TSOs, from third sector funding organisations, and from organisations who support small TSOs in conducting evaluation. Data from the interviews were analysed to identify themes relating to the use of programme theory in evaluation practice in small TSOs.

Key learnings

1. Participants reported very little use of programme theory in systematic evaluation activities. Despite this, there was a strong desire to use programme theory both internally within TSOs but also from funders who sought the kind of information about how and why programmes were working to support more targeted funding and to share learning about effective programme strategies with the sector. However, funders reported that it was difficult to develop reporting formats that include such information that are proportionate for small TSOs; that offer flexibility to TSOs to report on their work as they see fit; and that maintains some level of standardisation so that learning can be generated across reports.

2. Unsurprisingly, participants had strong implicit understanding of programme theory in terms of how they described their programmes. This understanding was likely supported by two things. First, the small size of the organisations meant that those conducting evaluation were also involved in service delivery. Close contact with programme implementation and beneficiaries facilitated the observation of change on a daily basis. Second, the professional training backgrounds of many participants, e.g. in social/youth work, also facilitated their understanding of need, change, and programme theory.

3. Whilst this implicit understanding of programme theory was not reported to be used within systematic evaluation activities, it was evident in more 'unsystematic' informal evaluation activities happening on a daily basis. 'Unsystematic' evaluation activities did not necessarily involve a formal evaluation design for data collection and analysis: rather it involved reflections by practitioners on a daily basis about how a programme was working and if not, addressing why this is so to improve it. The use of programme theory in this case is facilitated by good communication between staff members paired with the skills and opportunity to continually reflect on the programme.

4. Other uses of programme theory include the use of generic theoretical frameworks often supplied by government or local authorities e.g. asset-based frameworks or child-centred approaches. These kinds of frameworks guide programme development and implementation but are used at arms-length. The use of case studies also draws on programme theory however only at the level of the individual.

Implications

Whilst it is unsurprising that small TSOs have strong implicit understanding of programme theory and that this informs and guides their daily practices, it is important to acknowledge the desire to and potential value in the use of programme theory in more explicit and systematic ways. Of course, being mindful of constraints on time, resource and evaluation expertise is important but it should not be assumed that just because there is a lack of technical evaluation capacity that such organisations do not make use of programme theory. Rather, we must think of ways to support its use in more systematic ways that are of value to and proportionate for the evaluation needs of small TSOs, but also in ways that are useful to the other stakeholders who make use of such evaluation efforts.

Author: Robyn Millar , PhD student **Primary supervisor:** Professor Alec Morton

[1] Department of Management Science, University of Strathclyde Business School

References

Carman, J. G. (2011) 'Understanding Evaluation in Nonprofit Organizations', Public Performance & Management Review, 34(3), pp. 350–377. doi: 10.2753/PMR1530-9576340302.

Coryn, C. L. S. et al. (2016) 'Does Research on Evaluation Matter? Findings From a Survey of American Evaluation Association Members and Prominent Evaluation Theorists and Scholars', American Journal of Evaluation, 37(2), pp. 159–173. doi: 10.1177/1098214015611245.

Coryn, C. L. S. S. et al. (2011) 'A Systematic Review of Theory-Driven Evaluation Practice From 1990 to 2009', American Journal of Evaluation, 32(2), pp. 199–226. doi: 10.1177/1098214010389321.

Donaldson, S. (2007) Program Theory-Driven Evaluation Science: Strategies and Applications. New York: Lawrence Erlbaum Associates.

Ellis, J. and Gregory, T. (2009) Accountability and learning: Developing monitoring and evaluation in the third sector, Charities Evaluation Services. London.

Marchal, B. et al. (2012) 'Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research', Evaluation, 18(2), pp. 192–212. doi: 10.1177/1356389012442444.

Mark, M. M., Henry, G. T. and Julnes, G. (2000) Evaluation: An integrated framework for understanding, guiding and improving policies and programs. San Francisco: Jossey-Bass Inc.

Schwandt, T. A. (2003) '`Back to the Rough Ground!' Beyond Theory to Practice in Evaluation', Evaluation, 9(3), pp. 353–364. doi: 10.1177/1356389030093008.

Valentine, K. (2016) 'Complex Needs and Wicked Problems: How Social Disadvantage Became Multiple', Social Policy and Society, 15(2), pp. 237–249.

Understanding impact in complex environments: the benefits of using different strategies

GRAHAM THOM AND SARAH BROWN, SQW AND STEFANO CONTI, THE IMPROVEMENT ANALYTICS UNIT

This article is about what happened when two different organisations took different approaches to evaluating the same programme.

EVALUATING COMPLEXITY IN THE HEALTH SECTOR – CONFERENCE PRESENTATION SESSION SPONSORED BY THE HEALTH FOUNDATION

The programme: the Sutton Homes of Care Vanguard, comprised about twenty interventions aimed at improving the quality of care and patient health and wellbeing for residents of local care homes. The Vanguard was one of fifty sites awarded funding from NHS England (NHSE) to develop and test new models of care.

The evaluators: SQW Ltd, an independent research consultancy, was commissioned to do a process and impact evaluation of the Vanguard while the **Improvement Analytics Unit (IAU)**, a partnership between The Health Foundation and NHSE, was tasked with undertaking a comparative effectiveness analysis.

The key challenges encountered by the evaluators:

- **The environment was complex:** the population of both care homes and their residents fluctuated during the programme. Care homes opened and closed. Residents moved in and out of care homes or passed away.
- **The programme itself was complex:** it comprised multiple interventions and it was hard to identify when, where, and to what degree interventions were implemented in each care home.
- **There were many unobserved factors** that could not be accounted for such as the frailty of individual residents, the quality and intensity of delivery of the Vanguard interventions within recipient care homes, and the existence of other initiatives with similar aims

The approaches:

- SQW designed a theory of change for the Vanguard and gathered different types of evidence from different sources to examine the context in which the Vanguard was operating, the mechanisms that might be driving change and any outcomes. To analyse where the Vanguard had most effect and which intervention(s) was most likely to have influenced that effect they undertook a longitudinal comparison of care homes based on local hospital data, compared outcomes for care homes based on weighted scores for interventions received, and ranked care homes by performance. To further investigate the circumstances and nature of any changes, SQW surveyed friends and family and care home staff, interviewed friends and family, care home managers and the Vanguard and its partners, and ran focus groups with care home managers.
- The IAU set out to assess the impact of the Vanguard as a single intervention on selected hospital activity indicators for residents of Sutton's care homes. The IAU evaluation used a matched control group formed of new residents, aged 65 or above, of care homes from similar CCGs against which to compare outcomes for Sutton care home residents. The IAU had access to pseudonymised, individual patient-level data from the Secondary Uses Services on A&E attendances, potentially avoidable / emergency admissions, hospital bed days, falls and fractures, UTIs, total / out of hospital deaths for Sutton and comparator care home residents.

The results:

Evidence from both evaluations about the impact of the Vanguard was far from conclusive. SQW found reductions in A&E attendances, non-elective admissions and length of stay for residents of nursing homes. These reductions had some correlation with the weighted intervention scores for care homes. On the other hand, length of stay was the only indicator that decreased for residents of residential homes but there was no evidence of correlation with Vanguard activity.

	Nursing homes		Residential homes	
	Change	Intensity/ Attribution	Change	Intensity/ Attribution
999 calls	↑	Low	↑	Strong
A&E attendances	↓	Strong	↑	None
Non elective admissions	↓	Low	↑	trong
Length of stay	↓	Moderate	↓	None

One interpretation of these results is that the Vanguard started with, was more designed for, and had a longer relationship with nursing homes than residential homes.

Results from the IAU analysis were even more ambiguous (see table below): while no statistically significant impact on the chosen indicators could be detected, there was nonetheless some indication of an increase in potentially avoidable / emergency admissions among Sutton care home residents relative to the matched control group. This finding was precisely the opposite effect to what the Vanguard was aiming to achieve.

Outcome	Relative risk	95% CI
A&E attendances	1.00	(0.76, 1.32)
Emer. adm	1.52	(0.93, 2.50)
Pot. avoid. emer. adm	2.81	(0.90, 8.79)
Hospital bed days	0.89	(0.83, 2.48)

Reflections

Interpretation of evidence

Both the SQW and IAU evaluations were largely inconclusive, providing some indications about the likely impact of the Vanguard but without being able to draw any firm conclusions. It could reasonably have been hoped that two evaluations of the same programme would produce greater insight into any changes generated. In practice, given the challenges of interpreting the evidence in a single evaluation, doing so across two evaluations simply added another layer of complexity.

Client evaluation capacity

Both evaluations needed engagement and support from the Vanguard. The SQW evaluation had the benefit of working closely with the Vanguard team over nearly 3 years. This facilitated a better understanding of the details of the Vanguard's implementation and potential explanations for different findings, while providing an opportunity to explain the evaluation process and evidence to the Vanguard team. The IAU was challenged by a more detached relationship with the Vanguard while undertaking a more technically complex evaluation.

Susceptibility to evaluation

The design and implementation of the Vanguard offered exceptional challenges to both sets of evaluators. Some of the complex characteristics of the Vanguard, such as the fluctuating care home population, could not have been influenced by the Vanguard and would have presented a challenge to any evaluation. The difficulty of understanding these characteristics could have been partially mitigated by the availability of some or better data about them, for example about the frailty of care home residents.

However, the main problem was the lack of detail about how the Vanguard was implemented. There were multiple interventions but no control over which interventions were delivered in a home, when interventions were started or finished, no quality control of how interventions were implemented and no baseline of what the quality of care was in homes before they engaged with the Vanguard programme.

Thus, even when outcomes were identified, it was difficult to know what might have been responsible for them; whether the Vanguard, a particular Vanguard intervention or an entirely unobserved factor. One lesson from these evaluations is that an approach to programme implementation that is conducive to evaluation would be welcome by all evaluators, regardless of their approach.

Evaluating adaptive programmes: reflections from a mid term review (MTR)

ELBERETH DONOVAN, TEAM LEADER CROSS GOVERNMENT PROSPERITY FUND EVALUATION & LEARNING, WYG.

Over the past decade much thinking has been done on how development interventions should be altered to improve the impact and sustainability of reform. In 2014 the discussion on 'doing development differently' went mainstream, and attempts at adaptive management (AM) are now commonplace.

Adaptive management is an approach which enables the design, delivery, management and oversight of a project where there is a conscious and systemic effort to, at all levels and on an ongoing basis during the life of the project, learn about what works (and what doesn't), about changes in the political environment, and about the needs of local actors – and to adapt, in real time and informed by robust evidence, strategic aspects (including theories of change, results frameworks, approaches) and operational details (e.g. systems, tools, resources, budgets) with the aim of achieving systemic and sustainable change and increased impact.[21]

Although AM emphasises rigorous embedded monitoring[22], learning and reflection to inform programme delivery, donors simultaneously require external reviews of these programmes for both accountability and learning purposes. While such reviews can be undertaken through an ongoing Developmental Evaluation[23] approach (for example in the case of the DFID Zambia Accountability Programme) traditional summative evaluations are also used to consider if and how adaptive programmes have delivered results. Together with my colleagues, Dr Brendan Whitty and Eunica Aure, I shared some thoughts on evaluating this new breed of adaptive and politically smart programmes at the UKES conference in April 2019, drawing on our experience of evaluating the Accelerating Investment and Infrastructure in Nepal (AiiN) and the Strategic Support to the Ministry of Interior phase 2 (SSMI-2) programmes on behalf of the UK Department for International Development (DFID) in 2018.

AiiN is a seven year, £46.3M technical assistance programme aimed at addressing key constraints to inclusive and transformational growth in Nepal; the programme's three-pronged approach addresses investment in infrastructure, policy reform and financial instability and the programme was designed explicitly to be adaptive and politically smart.

21. Update from Donovan, E. and Manuel, C. (2017) Business Environment Reform Facility: Adaptive Programming and Business Environment Reform – Lessons for DFID Zimbabwe. DFID: London https://assets.publishing.service.gov.uk/media/5c7d4bf740f0b603d312114f/BERF-Adaptive-Programming-and-Business-Environment-Reform-in-Zimbabwe.pdf

22. Ramalingam, B. et al. (2019). Making adaptive rigour work: Principles and practices for strengthening monitoring, evaluation and learning for adaptive management. ODI: London https://www.odi.org/sites/odi.org.uk/files/resource-documents/12653.pdf

23. Patton, M. Q. (2011). Developmental Evaluation. London: The Guildford Press

This brief paper captures a few key reflections from the evaluation of AiiN, where the MTR aimed to **1)** identify lessons on the operationalisation of AM for DFID Nepal, **2)** make recommendations for programme delivery improvement, and **3)** verify result claims at outcome level to release payment tied to performance. In short, we were asked to explore if the programme contributed to outcomes, whether the programme was implemented in a politically smart and adaptive manner and if this approach contributed to the achievements of the results.

A key challenge in evaluating an adaptive programme is however that the programme should evolve over time, possibly quite drastically – and the outcome level results and the way in which support are delivered could change during the life of the programme and/or might not be clear until the programme is well underway.

A second challenge is that thinking around AM itself is still emerging, with principles and practices evolving. This meant that we had to think carefully about how we could make a judgement call on if and to what extent the programme was operating adaptively, without clear and agreed 'best practice' on AM available.

Our interest was however broader than simply knowing if and how the programme worked adaptively (i.e. if changes were made) – we wanted to understand if these changes were informed by rigorous evidence and a robust understanding of the political environment and local needs, if the adaptations were both technically sound and appropriate within the context, and if adaptation as a result of this interpretation and application of the evidence contributed to increased or more sustainable results. In other words, our challenge was to find a way to 1) evaluate contribution of a changing and emergent Theory of Change (ToC) and 2) make a judgement on process quality and the contribution of this process to the results claimed.

To do this we designed an evaluation which allowed us to understand what happened (impact evaluation) as well as how it happened (process evaluation). Using a theory-based approach, we combined a contribution analysis (CA) with an explicit process review of monitoring, evaluation & learning (MEL), political economy analysis (PEA) and management activities.

By applying Mayne's[24] six steps to contribution analysis we could make a credible causal claim about the contribution the programme was making to the observed results. This involved identifying three case studies (two of which required retrospective development of the ToC) and, due to resource

We designed an evaluation which allowed us to understand what happened (impact evaluation) as well as how it happened (process evaluation).

constraints, working with the implementers where necessary to gather evidence (including beyond existing M&E data), prepare the draft contribution story, identify and explore counterfactuals, and prepare the final contribution story. We also simultaneously identified indicators, based on the available literature, which we felt signified the use of an AM approach, and investigated how implementers undertook aspects such as 'real time' political economy analysis, ongoing monitoring and reflection, cross programme knowledge management and sharing, decisions around financial investment, and identification of outputs (i.e. which results they were prioritising).

We specifically explored through document analysis, focus groups and key informant interviews (KIIs) the process for generating, interpreting and using PEA and M&E data, and investigated how both DFID and implementers reacted to intelligence, evidence and findings. This included considering how decisions on which stakeholders to engage with (or not) were made; how local political experts were identified and recruited; the skills, knowledge and networks these local experts had access too; if and how programme and DFID verified, triangulated and/or quality assured PEA and local context specific intelligence; and how programme teams decided which information to prioritise, disregard and ultimately react to. We also reviewed the M&E plans, approaches, capacity and tools, and mapped and reviewed both DIFD and implementer decision making processes to understand how data were gathered, analysed and interpreted, and how decisions on which activities to explore, prioritise, scale up or down were made.

[24] Mayne, J. (2012) Contribution analysis: Coming of age? SAGE Evaluation accessed on 19 September 2019 https://journals.sagepub.com/doi/10.1177/1356389012451663

In addition to investigating the processes and the quality of the PEA and M&E data, we aimed to review the appropriateness of the decisions made based on the information available; this required comparison of context specific information generated by the programmes with that of the evaluation team (generated retrospectively) and a judgement on both the contextually relevant and the technical soundness of programming decisions.

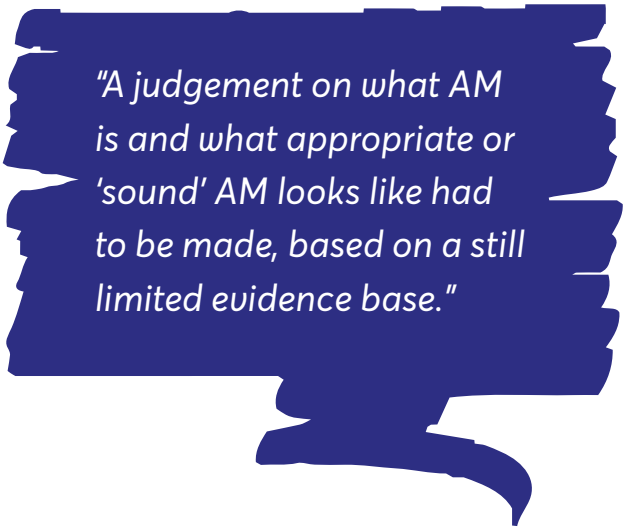
In short, our approach to evaluating the application of AM aimed to identify if the implementers had systems and processes in place which chimed with an AM approach (as captured in the literature) and which supported the collection of accurate information in the specific context in a rigorous manner. We also set out to understand if the implementers and DFID were analysing and interpreting this information appropriately, and if they were using it to make decisions which were both sound from a technical perspective and workable in Nepal- and which therefore improved impact and sustainability.

For the purpose of this article, I am not outlining the programme specific findings, nor the learning on how to 'do' AM based on AiiN as a case study; instead some reflections on the experience of undertaking an evaluation of an adaptive programme is summarised below:

- The evaluation was labour intensive and therefore expensive, requiring intensive in-country input from a dedicated AM evaluator, a local PEA expert and sector specific experts. The experienced local and international evaluators with strong sector expertise also required additional support and training on AM.
- The evaluation placed great demands (both timewise and intellectually) on the implementers; they had to play an active part in the evaluation to help unpack not only how they thought they worked, but how they actually worked, as well as reflecting on how they operationalised both concrete data gathering and decision making processes and undertook (often implicit) thinking and reflection processes.
- The evaluation involved evaluating the programme and the implementers' way of working as well as that of DFID, as the donor processes and decisions permitted, encouraged or hindered AM; this treatment of DFID as an active party to the programme raised its own challenges.
- A judgement on what AM is and what appropriate or 'sound' AM looks like had to be made, based on a still limited evidence base.
- The importance of having a well-documented and up to date ToC in place, which accurately reflects the hypothesis for change as it relates to what is actually being implemented is of paramount importance; the work required to reconstruct or update ToCs not only for the CA case studies but the programmes itself were significant.

As the implementation of programmes using an AM approach increases, so will the need to evaluate these programmes; this requires:

- **From implementers:** a greater emphasis on documenting and keeping up to date documentation on programme processes and decisions, as well as greater reflection not only on what is done but how and why it is done;
- **From monitoring experts**: an improved ability to adequately document ToCs, and to gather data on likely counterfactuals on an ongoing basis;
- **From donors:** a greater understanding of the active role they play in helping or hindering AM through the processes and systems they impose, and greater emphasis on ensuring adequate triangulation, interrogation and quality assurance of PEA and M&E information which informs programme decisions takes place;
- **From academics and think tanks:** greater support in documenting and sharing examples of AM in practice across the industry;
- **From evaluators:** the ability to take a 'systems approach' to evaluation and a requirement to improve the capacity of local evaluation experts able to engage with the generation and/or analysis of PEA data



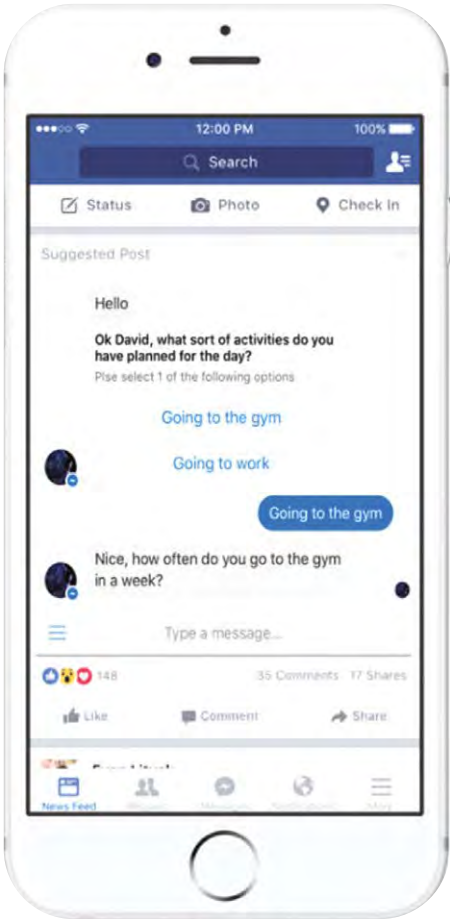
Digital in evaluation: a need to have not a nice to have

BEN COLLINS AND MARY SUFFIELD

EVALUATION DESIGNS, APPROACHES AND METHODS – A SPECTRUM OF INNOVATIVE PRACTICE – CONFERENCE PRESENTATION SESSION SPONSORED BY TRAVERSE LTD.

The use of digital in evaluation is rapidly becoming a need to have, not a nice to have. The benefits of using digital methods are multi-faceted, from improving data quality and research participants' experience to opening new frontiers of inquiry. Digital methods can improve the validity and reliability of the data upon which our evaluations are based. They offer improved recording and quality monitoring options and reduce the respondent burden by integrating into familiar platforms and allowing participants to respond at their convenience. Digital approaches also offer mechanisms for capturing observed rather than recalled behaviours, and avenues for accessing hard-to-reach groups.

From the many innovative technologies currently being used at Kantar, we have drawn out three to illustrate the usefulness of digital methods in evaluation: chatbots for 'in the moment' research; digital technologies to measure actual behaviour; and digital methods to achieve deliberative approaches.



Enabling 'in the moment' research

In many evaluations there is a time lag between when the experience takes place and when data is collected. Data can therefore be subject to recall bias, which reduces data validity.

Using a chatbot software that takes on a persona to conduct a conversation via textual or auditory methods as an interviewer offers the opportunity to overcome these problems and record experiences in the moment.

Chatbots can be deployed through a variety of platforms that participants may be familiar with, including Facebook Messenger. The interview begins as the participant is experiencing what we are measuring. We use natural language processing, which enables the chatbot to be interactive and responsive to the participant's answers. As a result, the interview becomes more relevant to the participant, putting them at ease and encouraging participation. We have found that chatbots increase response rates and reduce dropout rates. Chatbots can also be more cost-effective than other interviewing methods as the approach reduces the need for interviewer time and travel.

Case study: Evaluating a sexual health communications campaign

In the UK, Kantar has used a chatbot for a study to evaluate a communications campaign on sexual health among young people (aged 16-24). Targeted advertisements were put on Facebook and Instagram, which then directed participants directly to the chatbot which was hosted on Facebook Messenger. The intention was to give real time results on recognition and reaction to the campaign.

The use of a chatbot was a success. The number of completed responses was reached more quickly than when using a standard online survey and users' feedback was very positive. A number of factors contributed to the positive outcome:

- The chatbot offered young people the opportunity to participate in the survey in an environment that they spend a lot of time. In this way, it was less artificial than a classical online survey;
- The survey was better targeted at individuals who were more likely to have seen the campaign;
- The survey discussed sensitive issues that participant's may have been less comfortable discussing with a human interviewer.

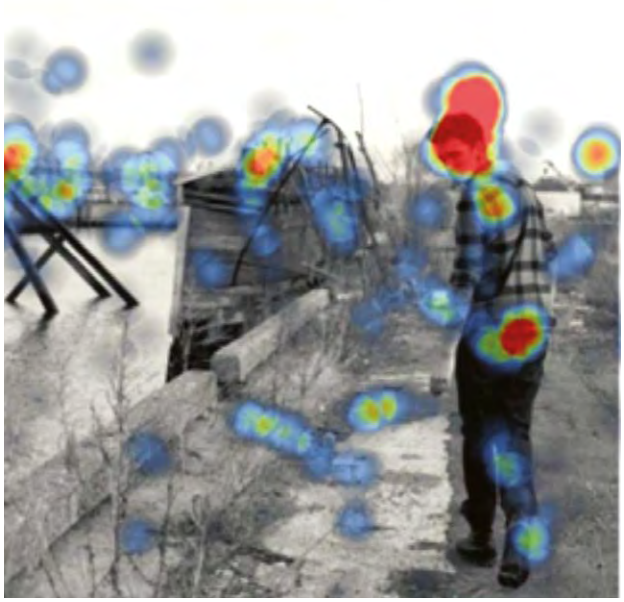
Measuring actual behaviour not just perceptions

Policies, programmes, and campaigns often seek to influence behaviours, and necessitate measuring changes in behaviours to evaluate their effectiveness and impact. Classical methods of data collection, such as surveys and interviews, often measure perceptions and recalled behaviours rather than actual behaviour. Participants may not remember exactly how they behaved in a situation or may provide what they consider to be socially acceptable responses.

These are both challenges to data validity and can lead to false conclusions. Digital methods, such as sensors and eye-tracking, can be used to more accurately capture behaviour.

Sensors can measure how frequently (or for how long) something is being used. For example, installed sensors can measure how frequently children are washing their hands, rather than relying on reported hand washing. They can be used before an intervention to measure existing behaviour and after to measure whether there has been any meaningful change over time.

Eye-tracking technology has been successfully used when evaluating communications campaigns. It can identify which elements of the campaign are attracting the most attention and can code the participant's facial expression to understand whether the response is positive or negative.



Case study: Evaluating a behaviour change intervention in Ghana

Pneumonia and diarrhoeal diseases are one of the major causes of death for children under five in Ghana. Handwashing with water and soap (HWWS) is a proven intervention to reduce this rate. However, very few Ghanaian households' practice HWWS.

In collaboration with UNICEF, Kantar conducted an evaluation to measure the effectiveness of an intervention to change hand washing behaviours in Ghanaian schools. Sensor hardware was sealed in a waterproof case and positioned on jerrycans. The sensors measured the number of handwashing events per hour before and after the intervention. The data was used to explore the extent to which there had been an increase in handwashing events since the intervention.

The evaluation showed that children in the intervention group practised HWWS significantly more than the comparison group, demonstrating a positive impact of the intervention.

Digital deliberative dialogues for policy formation and understanding

At the heart of any intervention – whether it is a programme or campaign – is policy. To effectively evaluate, we must first understand not just what people's thoughts and beliefs are, but also how they have formed them, and the factors that shape those views. Understanding thought processes allows us to design, communicate and improve policy. It can also improve our understanding of a policy and help us to more deeply evaluate it.

Deliberative approaches allow researchers to effectively understand both a participant's beliefs and the thought processes behind them. It is a distinctive approach to public engagement that provides people with information and gives them the time and space to discuss and deliberate on an issue in depth, before coming to a considered view.

Digital deliberative dialogues use research-facilitated online communities as a platform for these conversations to take place. They can host participants, researchers and experts on the given topic, and include a range of activities to bring ideas to life. Digital dialogues are flexible. They are typically open across multiple days and can be accessed by mobile, desktop and tablet. Participants can dip in and out when it suits their diaries. They also allow participants to be more open than with other methods because of the anonymity that being online offers. Finally, deliberative dialogues are more cost-effective than the face-to-face alternative.

Case study: Digital deliberative dialogue on future roads

In the UK, Kantar hosted a digital deliberative dialogue based on scenarios simulating mobility services in 2030 -2050. The work was commissioned to better understand public responses to imminent technologies such as electric and automated vehicles, and what information is needed to be communicated to make the public more comfortable with their use.

Videos, animations and games were used to immerse participants in the scenarios, understand how they were forming their views and ultimately gain a detailed understanding of their considered view.

A comparison was drawn with the same dialogues that were held face-to-face and the content and quality of the outcomes was consistent.

Taking digital forward…

This article provides a snapshot of just three of a much broader range of digital methods being employed by Kantar to meet our clients' needs. These build on evaluation principles and classical research methods to enhance data quality and validity and provide new measurement options that were previously infeasible. Evaluators must continue to explore how digital technologies can be brought to bear on our work with clients to increase the impact of their social policies and interventions.

Practicing data visualisation as an evaluator

DAVID DRABBLE, SENIOR RESEARCHER, TAVISTOCK INSTITUTE

EVALUATION DESIGNS, APPROACHES AND METHODS – A SPECTRUM OF INNOVATIVE PRACTICE – CONFERENCE PRESENTATION SESSION SPONSORED BY TRAVERSE LTD.

TRAVERSE

At certain points, all evaluators will find it difficult to communicate their findings to their audiences. Hiding key messages in 100 page reports, lost in a swamp of figures, visually highlighting important findings can be a lifesaver in explaining what you found. One method of improving how you communicate your findings is to improve your data visualisation skills. Your findings are your currency, and when they are fully understood it increases the value of your conclusions and recommendations.

Effective data visualisation is about telling a convincing story. In his book Data Visualisation: A Handbook for Data Driven Design (2016), Andy Kirk defined data visualisation as ***"The representation and presentation of data to facilitate understanding"***. This is a useful definition in highlighting four components. All data visualisation needs to have underlying data, with carefully selected charts, attention to style and presentation, and the intention to lead readers along a path to understanding.

Given the various considerations that need to be taken into account, staging how you build a visualisation can help make the process easier to execute.

Stage 1: Data cleaning and pattern detection

Whether you have qualitative or quantitative data, it is important to first develop your data before any visualisation, cleaning it to make it easier to categorise. After doing this, **look for patterns and surprises** in the data. It can help to print out the data and read it – this can be done even with a large Excel sheet. After looking for patterns with an open mind, relate these findings to your e**valuation questions** – answering these questions is the purpose of data visualisation for an evaluator.

Stage 2: Development of a narrative

The evaluator is a **sense-maker** so you shouldn't present everything you find. As a communicator, writing up all your findings with charts of every survey question is not illuminating, it is confusing. You should consider your narrative and **story board** what you want to say with your data. What is the hierarchy of information you want to present? What's most important and how will you explain this finding? Staging a story helps you think through the order of information you will present.

Stage 3: Choosing your charts

In selecting which charts to use, first consider the type of data you have alongside the story you're trying to tell. There are some shortcuts in making these decisions – the FT's Visual Vocabulary is particularly helpful in understanding your **options for representing different types of quantitative data**. In general, if you have more than four pieces of data, don't use pie charts, if you have time based data, use lines or bars, and if you have location data, consider using a map. As a rule, bar charts are your friend as they have few drawbacks in interpretation and are flexible.

Stage 4: Iterations of design

Design is the final stage and often the most difficult for evaluators. It can be tempting to use default charts and colours from Excel but to **create something memorable** and fit to your specific data and story it is always helpful to engage in design. This can be daunting but don't be discouraged about your expertise and work within your limits. You can use shortcuts if you aren't a designer – it helps to using data analysis software that has well designed defaults such as Tableau Public. If you spend a few days learning it, you can also use **design software** to clean up the graphics: the basics of Adobe Illustrator are possible to quickly pick up. If not, PowerPoint is a good fall back option.

These four stages imply that data visualisation needs a wide variety of skills: data collection skills, analytical skills, storytelling skills, and design skills, in terms of aesthetics, user experience and the technical skills to use software. Rather than suggesting that all evaluators become Leonardo Da Vinci, it is more practical to **work with people from different disciplinary backgrounds**.

Making an impact

Once you have a satisfactory set of graphics that represent the narrative you've developed, there are a number of ways you can use these figures to achieve an impact with the client. Ask the client what findings would be useful to hear about as **early as possible** – your interest and enthusiasm aren't always going to be reciprocated! You can use workshops and meetings to **drip feed** your visualisations, and I'd always recommend not saving everything for the final report. You can use feedback as a means to revise and improve your work in this way. Ultimately, **data visualisation is a communications device** and an effective one. Findings that you think will be hard to accept can be made easier to swallow by convincing data visualisation as these show how you came to your conclusions clearly.

If possible, try to share these visuals outside the client system – if you have worked through these four stages **your figures will be products** rather than just graphs, so use them as such. Social media is a very good way of sharing infographics as you will get more engagements with well-designed figures compared to ordinary news items. Sharing on your website is also useful for an impactful summary of evaluations. When you share graphics with a wider audience try to be loose and more **playful** if possible to engage others. In the end, doing this well, and getting an outside audience to engage in our evaluations through data visualisation **can make important work live on**.

Using realist evaluations to design and implement complex interventions: Experiences from a market-systems development programme in Ethiopia

MATTHEW MCCONNACHIE AND CLARISSA SAMSON,
LTS INTERNATIONAL

Our UKES presentation described our experience from drawing from a realist-evaluation approach to inform how an Ethiopian development project was designed and implemented. In this article we first introduce the realist approach and the project before describing our experiences and lessons.

Background to the realist approach and its suitability for designing interventions

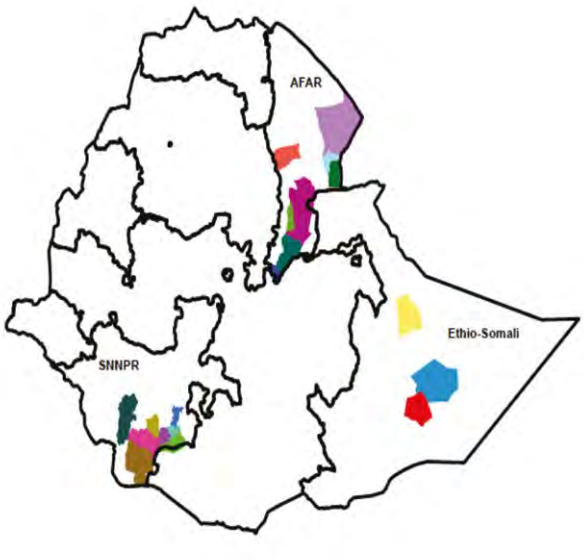
Realist evaluations are focussed on understanding the mechanisms which underpin interventions and how they interact with specific contexts. This makes the realist approach well-suited for extrapolating findings to new contexts and informing how interventions are designed and implemented. The approach can provide findings that are linked to specific contexts and target groups, and explain the generative mechanisms, explaining why change happened. Despite the high-potential, limited work has been done on how to use realist evaluations to inform the planning and design of interventions

Case study background: The Market Approaches to Climate Resilience in Ethiopia project

The project that we evaluated was part of the DFID programme called: "Building Resilience and Adaptation to Climate Extremes and Disasters programme (2015 – 2019)" (BRACED). It was implemented by two NGOs: Farm Africa and Mercy Corps. The project aimed to strengthen the private-sector and community-management systems for lowland Ethiopians (financial services, natural resource management, urban business opportunities). The project aimed to bring about transformational change using a systems-based approach and working across multiple levels (from individuals to local and national institutions).

All projects within the BRACED programme were required to use the realist evaluation approach, a realist synthesis was used at the programme level. LTS did the evaluation work on behalf of the project implementers. Evaluation deliverables included periodic reports (M&E plan, baseline, mid-term, endline and extension evaluation reports).

FIGURE 1. The photograph shows the Afar region in Ethiopia of pastoralist farmers. The map shows the project locations in Ethiopia (coloured polygons show the Woreda (district) outlines).



Our experiences and key lessons of using the realist approach for project design and implementation

Below we briefly outline our experiences of using the realist approach to inform the design of the project during the early planning and implementation stages. We end by outlining our key lessons.

Experiences during the initial design-phase

The evaluation team played an active role during the project proposal and planning stages. This helped to ensure good alignment between the evaluation and implementation work, especially for deciding on the indicators and targets for the project and drawing from findings from previous studies to inform project design. With hindsight, however we would have benefited from drawing more strongly from the realist approach during the initial proposal and design stages. For example, we identified evidence gaps but not entirely through a realist lens (e.g. evidence gaps related to specific contexts and sub-groups). We would have also benefited from developing our Theory of Change in a more explicit realist format during the initial project design stage.

Experiences during the implementation phase

We found that the realist approach provided useful insights for adapting the design of the project at the mid-term and endline into the extension phase. We used data collection templates that were context and target group specific, project implementers found it useful to have the evaluation findings structured around specific target groups and contexts. Based on feedback from the BRACED Knowledge Manager, we generated some useful findings for the BRACED programme-level evaluation.

Despite the above advantages of using the realist approach, we encountered some challenges. We could have benefited from more ongoing real-time assessments, tightly linked with the monitoring work done by the project staff. This could have helped to inform project design on a real-time basis and not only at the mid and extension stages. We also found it challenging to sequence the collection of quantitative (household survey) and qualitative (focus groups) data so that the one could inform and build on the other. In realist evaluations the quantitative data is more useful for understanding what changes happened whilst the qualitative data can help with understanding how and why the changes happened.

Another challenge we encountered was finding the balance between providing project-wide generalisable findings and realist case study deep-dives. The focus of realist evaluations is to provide findings linked to specific contexts and sub-groups. This focus is well suited for learning purposes but less so for accountability purposes where knowing the average effect of the project is important.

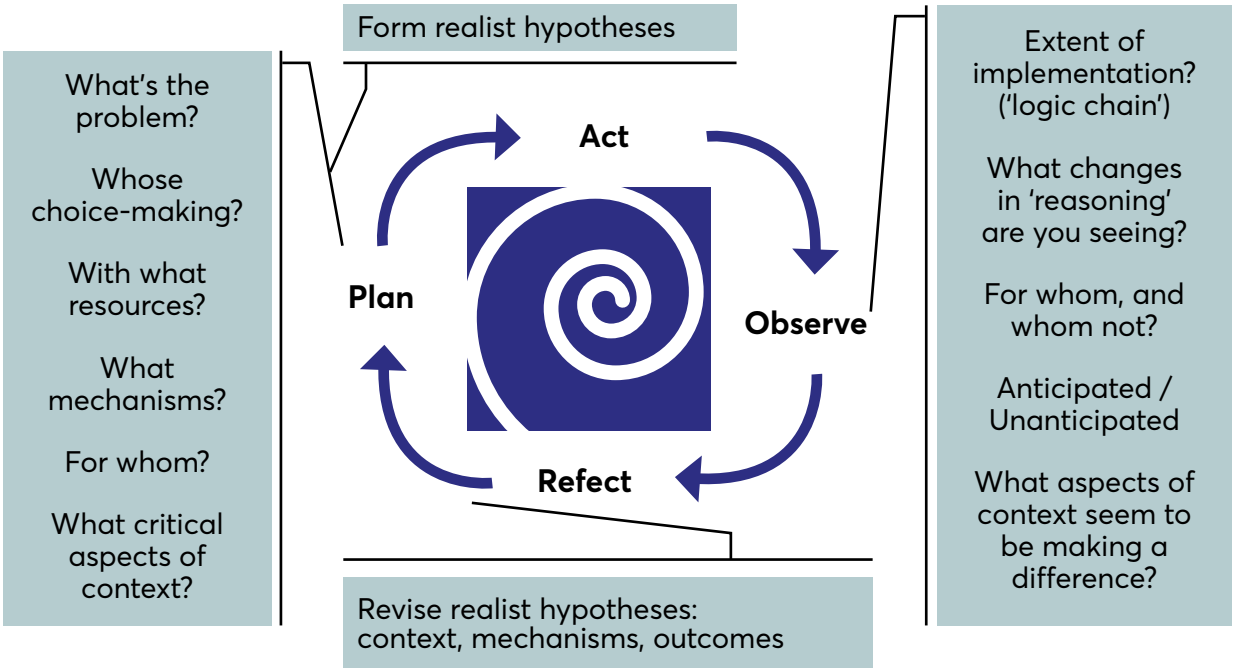
The key lessons that we learnt:

- Integrate the realist approach with planning-focused frameworks (e.g. co-design, adaptive management, developmental evaluation). This helps to provide an operational framework for linking the evaluation findings with implementation design. Box 1 below provides an example of how this could be done).
- It is important to develop the realist program theory concurrently with the design of the program, iterating back and forwards between them. This requires tight links between the evaluation and implementation teams, which can be assisted by using a planning-framework (above lesson).
- Identify the learning-gaps before designing the project. The realist approach can be resource-intensive, the realist evaluation work should be prioritised to where the evidence-gaps are.
- Establish adequate processes and systems (workplans, adequate resources). Linked to the first lesson, the realist and planning frameworks need to be effectively and efficiently project managed.
- Be participatory and provide tailored communication of findings to the evaluation audiences. Realist evaluation methods can be heavy on technical terms and jargon. A crucial aspect is to work in a participatory way with project implementation staff and other evaluation stakeholders.
- Situational knowledge will always be important. Given the complex nature of development projects like our project, evaluation findings will always need to be supplemented with the situational knowledge of project implementation staff.

"The evaluation team played an active role during the project proposal and planning stages. This helped to ensure good alignment between the evaluation and implementation work."

Box 1: Example of linking the realist approach with planning-frameworks

Westhorp, Stevens and Kaye (2016) used realist action research in 'The Bridgewater Project', to evaluate different social security payment options. Action research seeks to solve real world problems, trialling solutions until a 'best fit' solution is reached. The authors also integrated this with a co-design approach to ensure that a collaborative approach to service design was used. The main benefit they found was that it helped project staff make sense of findings and identifying the needs of beneficiary sub-groups.



Acknowledgements

We would like to thank the project staff from Farm Africa and Mercy Corps for their assistance with the evaluation and the important lessons that we learnt from them. We would also like to thank Gill Westhorp, Rebecca Hardwick and Jake Lomax for their ideas and feedback on our evaluation.

On the art of robust generalisation: reflections from field-testing the qualitative impact protocol (QuIP)

JAMES COPESTAKE

Impact evaluation goes beyond credibly demonstrating that intervention X contributed to outcome Y in context Z; it also concerns the scope for robust generalisation about how far such causal links are likely to apply to a predictable range of other contexts. For realist evaluators, the term robust generalisation connects both to the idea of middle range theory and to the multiple Context-Mechanism-Outcome (CMO) configurations that help us to understand not just the average effect of X on Y across a specific set of contexts, but how it varies across them and why, so that we are better informed about what is likely to happen in a new situation. In other traditions, the phrase also connects with the idea of theories of change and claims to external validity.

This article reflects on five years of action research into impact evaluation based on narrative data, relying on self-reported attribution, and the latent counterfactuals within everyday speech, rather than inferring attribution through statistical analysis of variation in outcomes across a populations subject to variable exposure to an intervention. More specifically, it reflects on the how this approach can contribute to robust generalisation. We first spent three years designing and testing a qualitative impact protocol (the QuIP) for assessing the impact of specific rural livelihood strengthening projects in Ethiopia and Malawi. That done, the awkward question remained - how robust was the QuIP (itself a set of methodological generalisations) for use in other contexts? This prompted two more years of action research, during which we used the QuIP to carry out commissioned impact evaluation studies in a range of other contexts. This work is reported in Attributing Development Impact: the qualitative impact protocol case book (Copestake, Morsink & Remnant, 2019), available free as an e-book at **bit.ly/QuIP-OA**.

Box 1 provides a brief description of the QuIP, culled from the book, which also includes the full protocol in an Annex. The reason for setting this out explicitly was to promote transparency, and confront the view that uncertainty over how qualitative impact evaluations generates useful evidence reduces demand for it. The QuIP was of course adapted and applied differently in each of the ten case studies of its use reported in the book: seven of the studies were 'double' QuIPs, for example (see point 4 in Box 1). Some studies also set out primarily to confirm explicit theories of change, while others were more exploratory. Further differences in commissioners' goals, details of data collection and analysis are summarised in Chapter 10 of the book. Here I focus on what we learnt about how commissioners actually used the information generated.

BOX 1. A BRIEF DESCRIPTION OF THE QUIP

1. The QuIP is a standardised approach to generating feedback about causes of change in people's lives that relies on the testimony of a sample of the intended beneficiaries of a specified activity or project.
2. The scope of a study is jointly determined by an evaluator and a commissioner, the shared purpose being to provide a useful 'reality check' on the commissioners' prior understanding of the impact of a specified activity or set of activities.
3. A single QuIP is based on the data that two experienced field researchers can collect in around a week. A useful benchmark (that emerged through the design and testing phase) is that a 'single QuIP' comprises 24 semi-structured interviews and four focus groups. Specific studies may be based on multiples or variants of this.
4. Interviewees are selected purposively from a known population of intended beneficiaries, ideally after analysis of what available monitoring data reveals about the changes they are experiencing.
5. Where possible initial interviews and focus groups are conducted by independent field researchers with restricted knowledge of the activity being evaluated. This means that respondents are also unaware of what intervention is being evaluated, a feature referred to as double blindfolding.
6. Transcripts of interviews and focus groups are written-up in pre-formatted spreadsheets to facilitate coding and thematic analysis.
7. An analyst (not one of the field researchers) codes the data in several predetermined ways. Exploratory coding identifies different drivers and outcomes of change (positive and negative). Confirmatory coding classifies causal claims according to whether they explicitly link outcomes to specified activities, do so in ways that are implicitly consistent with the commissioners' theory of change, or are incidental to it.
8. Semi-automated generation of summary tables and visualisations of what the coding reveals to aid interpretation of the evidence.
9. It is easy to check back from summary evidence to raw data for purposes of quality assurance, auditing, peer review and deeper learning.
10. Summary reports of the evidence are a starting point for dialogue and sensemaking between researchers, commissioners and other stakeholders thereby influencing follow-on activities.

Box 2 provides a summary how commissioners made use of the QuIP studies, based on semi-structured interviews with them between one and two years after the study was completed. This included promoting reciprocal learning through follow-up stakeholder workshops, providing evidence to inform operational decisions, and dissemination of findings to the wider public. Five of the organisations also went on to commission further QuIP studies.

BOX 2. REPORTED USE OF SELECTED QUIP STUDIES

Case study	Promoting reciprocal learning?	Evidence for operational decisions?	Shared with the public?	Further QuIP stud-ies?
Diageo Ltd; purchasing of malt barley from small-scale farmers; Ethiopia.	No	Yes	Yes	Yes
C&A Foundation; empowerment training for garment workers; Mexico.	No	Yes	Yes	No
Terwilliger Center; refinancing and training for housing microfinance; India.	No	Yes	Yes	Yes
Tearfund; church and community mobilisa-tion; Uganda.	Yes	Yes	Yes	Yes
Save the Children; harnessing agriculture for nutritional outcomes; Tanzania.	Yes	Yes	No	Yes
GSHP; placement of volunteer health edu-cators; Malawi, Tanzania & Uganda.	No	Yes	No	No
Frome Town Council; promoting use of green urban spaces; England.	No	Yes	No	No
Oxfam; fairtrade coffee and women's em-powerment; Ethiopia.	No	No	Yes	No
Acumen; impact investment (dairy and beauty parlour franchise); India.	No	Yes	No	No
Self Help Africa; integrated area development; Zambia.	No	Yes	No	Yes

Promoting reciprocal learning through stakeholder workshops. Tearfund took findings back to the villages where interviewing took place, while Save the Children staff organised stakeholder meetings both in the project area and Dar Es Salaam. Other commissioners cited lack of funding for not organising follow-up activities, but their reluctance in some cases also revealed the priority attached to hitting measurable programme targets and controlling costs over longer-term and less tangible goals, including empowering partners.

Evidence for operational decisions. Most respondents reported that evidence from the studies had indeed influenced operational and funding decisions, but they mostly opted to do so anonymously. Establishing how far evidence influences action or not is itself methodologically challenging. It can also be politically sensitive given the often wide gap between organisational goals and ground realities. Anticipating this, an important lesson is that potential users and producers need to formulate a clear and early understanding of what to expect from an evaluation study as a form of political deliberation. How open to challenge and to change are they?

Dissemination to the wider public. Five out of ten of the commissioners listed in Box 2 had published full or abbreviated versions of the QuIP reports. These faithfully covered negative as well as (mainly) positive findings, with editing focused on shortening, simplifying and polishing presentation and ensuring the descriptive text was consistent with other material published by the organisation.

An unexpected finding that emerged from writing the book is that the primary motivation for commissioning independent QuIP studies was often not to produce timely feedback to inform operational decision-making. Rather it was to inform a longer-term process of reflecting upon and defending the robustness of middle-range theory about what the commissioning organisation was aiming to do and how.

Box 3 illustrates this by listing some of the core generalisations addressed by these studies.

BOX 3. SUGGESTED MIDDLE RANGE THEORY BEHIND SELECTED COMMISSIONERS' ACTIVITIES

Case study	Suggested theory
Diageo; malt barley promotion; Ethiopia.	Purchasing malt barley as a cash crop from small-scale farmers does not have unintended negative social consequences.
C&A Foundation; garment worker training; Mexico.	Garment factories can offer their employees 'empowerment' training that improves both their relational wellbeing and productivity.
Terwilliger Center; housing microfinance; India.	Incremental home improvement funded by commercially self-sustainable housing microcredit benefits borrowers and their households.
Tearfund; church and community mobilisa-tion; Uganda.	Faith-based community development can have a positive transformative effect, even when not linked to material transfers.
Save the Children; harnessing agriculture for nutritional outcomes; Tanzania.	Important synergies arise from integrating agriculture, nutrition and gender training activities together, rather than intervening in each area separately.
Global seed health partnership; Malawi, Tanzania & Uganda.	American health care volunteer educators can make a positive contribution to the training of doctors, nurses and midwives.
Frome Town Council; promoting use of green spaces; England.	Council supported amenities and events can have a positive effect on citizens' wellbeing by influencing the way they use parks and other green spaces in the town.
Oxfam; producing fairtrade coffee; Ethiopia.	Promoting fair trade coffee as a cash crop does not have adverse effects on the wellbeing of women by increasing their work burdens.
Acumen; impact in-vestment; India.	Blinded telephone interviews with company clients can provide useful feedback of the social impact of their services.
Self Help Africa; integrated area develop-ment; Zambia.	Integrated interventions to promote agriculture, nutrition, health and social relations can have a transformational social impact – more so than intervening in each area separately.

The distinction between reflecting upon and defending robust generalisation connects with the point about political deliberation. The middle range theory embedded in the core generalisations of organisations are bound up with their identity, vision and mission. It follows that it is a matter of politics how far, and how openly, they are willing to reflect critically about them. This highlights the importance of case selection to robust generalisation. Politically safer case selection seeks out contexts where the likelihood of confirming core generalisations is greatest, or where context diverges so radically from the normal that zero or negative impact can be ignored. In contrast, the most interesting but also politically risky cases are those at the margins of confidence in our generalisation. This highlights the importance of including case selection in political deliberation over study design, as well as why case selection should be as fully informed as possible by knowledge of (a) the organisation's existing core generalisations and (b) relevant contextual variation across its area of operation. This view of case selection has a Bayesian flavour quite different to that of classical statistical sampling theory. It also differs from the logic of case selection to achieve saturation associated with more inductive, exploratory and theory generating social research.

How well do the ten case studies themselves hold up at a test of the QuIP's usefulness as a set of methodological generalisations? While they were mostly selected opportunistically in response to approaches from commissioners they did enable us to explore their usefulness in a wide range of new contexts. The results have been sufficiently encouraging to encourage us to continue to explore for whom, where, when, how and why QuIP and other qualitative impact evaluation studies can be useful.

The Measurement of Diversity

RICK DAVIES

No, this is not yet another example of Obsessive Measurement Disorder? In this paper I want to outline a number of arguments about why and how we should think about measuring diversity.

There are two broad reasons for thinking about the measurement of diversity. The first is normative, because we think diversity is important. For example, we want to see people from all sorts of backgrounds to have equal opportunities and be treated equally and fairly. The second is pragmatic, because we think the existence of sufficient diversity can make a practical difference to how we do things and what we achieve. To some extent, the first of these is about the outcomes we want to achieve whereas the second purpose is about means, how we do things.

Diversity is a multidimensional concept. As you might expect there is a big literature on the measurement of diversity in the field of ecology. See for example the Wikipedia entry on the Measurement of Biodiversity. There you will see distinctions made between species diversity, ecological diversity and genetic diversity. My interest is in how we can take measurement approaches from these fields and make them useful for evaluation purposes when we are talking about human societies. Why do so? Well, I am inclined to agree with this quote: "Ninety per cent of problems have already been solved in some other field. You just have to find them." (Tony McCaffrey, 2015).

Traditional biodiversity measures have combined two attributes: the number of types of things such as species, and the relative numbers of each of these types. These two attributes are known as richness and evenness. There is a third attribute which is about phylogenetic diversity; this is the measure of the amount of difference between the types. For example, there is a bigger genetic difference between people and frogs than between people and chimpanzees. In a seminal synthesis paper Sterling (2007) has labelled these three measures as variety, balance and disparity. The first two of these three measures seem fairly easy to transfer to human settings. But the third one looks more problematic because many important human differences are not measurable in genetic terms. But there are ways around this as I will explain below.

The pragmatic justification for paying attention to the measurement of diversity is that different forms of diversity can have different consequences for what groups of people can achieve. A certain degree of homogeneity is necessary for people to be able to communicate effectively, in that respect we can see some forms or degrees of diversity as problematic. There are clearly some minimal requirements. On the other hand, it is argued that in some settings greater diversity can make groups more effective than less diverse groups. One interesting question here is in what sort of settings is increased diversity useful? Scott Page (2017) has argued that it is where work is cognitive rather than manual and nonroutine rather than routine. Particularly where tasks are not easily decomposable. Another question is what types of diversity are most useful in such settings. Can different forms of diversity be categorised? Page makes a distinction between cognitive diversity and identity diversity. He argues that it is the former which makes the biggest difference to performance of groups, but that identity diversity can enhance cognitive diversity. Note that this is a separate argument to where the concern is about the normative value of diversity, where we might expect identity diversity to be prioritised.

I now want to talk about one application of these ideas. Over the last six months I have been working on the development of web app called ParEvo. ParEvo is web-assisted participatory scenario planning process. Another way of describing it is as a form of participatory futures research. Futures research has been defined as '… a transdisciplinary field of inquiry that uses a variety of methods to explore possible, plausible, and preferable futures. The goal is to develop foresight-insight into how and why the future could be different than today-to improve policy, planning, and decision making" (Bengston, 2019). This is an area where the diversity of participants could be expected to make quite a difference to the results that are generated i.e. the types of futures that are envisaged. During the ParEvo development phase I ran two scenario planning exercises involving 23 people from 13 different countries. Participants varied in the extent to which they built on each other's contributions and the extent to which others built on their contributions. The different scenarios, emerging as storylines, varied in the extent to which they were built by a wide range of participants versus only one or two. As the storylines developed some died out and others develop multiple new branches. Some were close relatives, others for more distant. All these behaviours could be described using the three different diversity measures mentioned above. The outcomes of interest are the different types of scenarios that are generated by this participatory process. So far, these are evaluated by the participants at the end of the process on two criteria: their possibility of occurring in real life and their desirability. It is also possible to introduce other evaluation criteria, such as verifiability. At this stage I have no specific prior theory of what types of participant behaviour will lead to what types of outcomes. But a structure now exists for exploring these issues.

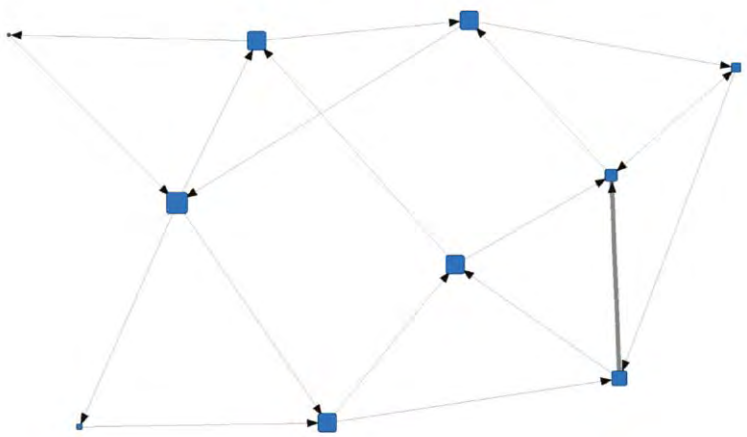
Figure 1 shows how different the behaviour of participants in different scenario planning exercises can be. In the upper example participants built on a range of other participants contributions. Nodes are participants and arrows show which participant built on which other participants' contributions. Thicker lines show more contributions. There is a clear difference in the behaviour of participants in the two exercises. The lower example was a scenario planning exercise about possible Brexit developments, post March 2019.

FIGURE 1

DISPARITY MEASURES

NETWORK DENSITY = 37%

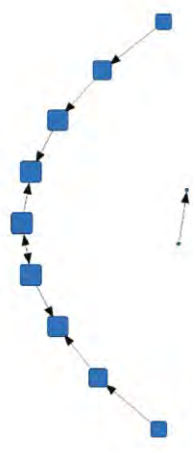
AVERAGE "CLOSENESS" = 2.4



DISPARITY MEASURES

NETWORK DENSITY = 23%

AVERAGE "CLOSENESS" = 6.6



In these types of exercises the disparity dimension of diversity can be measured by borrowing some ideas from social network analysis. In social network analysis distance between actors is measured in terms of the numbers of links necessary to traverse between one and the other. 'Closeness', the shortest distance between any two actors, is one measure of distance. In Figure 1 we can see that the average closeness between the actors in the first exercise is much less than the average closeness in the second exercise. Another social network analysis measure is network density, which in simplest terms means the proportion of all the possible connections in a network that are actually present in the network. The more connections that are present the closer the actors are likely to be to each other. In Figure 1 we can see that there is a big difference between the two groups in their network density. The counter-intuitive idea of measuring disparity by asking what people, things or ideas are most similar, is more widely generalisable if the results are analysed in terms of network structure. For example, during evaluation fieldwork one can easily ask about perceptions of similarity. Such as those between programme participants, or between income sources, etc.

Going back to Page's distinction between cognitive diversity and identity diversity the results we can see here probably reflect cognitive diversity more than identity diversity. Participants' identities were not disclosed or attached in any way to the contributions that were being made. So, when viewing Figure 1, the individual nodes should be seen as sets of ideas, each with their own common source.

This sort of structured experimentation with participatory production of knowledge does not need to be analysed in isolation. In the field of social psychology there is already a significant body of knowledge about how group structures affect task performance. More recently, a field known as collective intelligence has been specifically concerned with the design of groups that can perform better than their most capable member. In the context of ParEvo exercises one proposition that could be tested is whether storylines constructed by multiple participants perform better on criteria such as verifiability than those constructed by individuals. And more particularly, if more balanced contributions from the multiple participants makes any difference.

Building knowledge about what works more effectively is not the only reason why we should be measuring diversity. Going back to the idea of the normative value of diversity, there are plenty of circumstances where in the course of an evaluation one might want to measure the extent to which value-based expectations of inclusion have been achieved. Thinking of inclusion from the diversity perspective would make sense. Diversity is measurable in a way that is not narrow and reductionist. Rather, the combination of 3 measures introduced above can help us appreciate the range of what is possible.

Bringing out the best in times of uncertainty: re-evaluating collaborative evaluation roles

GEORGIE PARRY-CROOKE, PRINCIPAL RESEARCHER/ CONSULTANT, TAVISTOCK INSTITUTE

Georgie Parry-Crooke, Principal Researcher/Consultant at the Tavistock Institute of Human Relations and Professor Emerita of evaluation and social research at London Metropolitan University, has always been interested in evaluators and their roles which she continues to explore here.

The role of the evaluator is rarely static and debates about roles shift along a continuum from the supposedly objective to neutral to independent while making evidenced judgments through the use of varied methodologies and methods. Times also change and it may be that there are more opportunities now to speak about who we, the evaluators, are and how we work. In addition, the current period of financial, political and social uncertainty may be forcing evaluators and evaluation commissioners to pay closer attention to role now. We wear many hats (and sadly no space to include the ones I chose for the conference!).

The relevance of role is contextual. It is the context and our increased understanding of complexity - doing complex evaluations and doing evaluations in complex environments - which help to define our roles. There can be increased pressure to reassure organisations, programmes, projects and people experiencing increasing uncertainty, giving a veneer of certainty, where it cannot be provided. And perhaps this offers opportunities to open up about who we are through consideration of different ways in which to collaborate. This may lead to shared understandings of what is happening even if they are not always comfortable to live with.

Evaluators can reconsider the question of who we think we are and where we stand in relation to evaluation practice. This may be inside or outside organisations involving skills and competence, independence yet developing 'friendship' and experiencing anxieties while we work within settings where often funding is shrinking. And while some wish evaluators to give the impression of certainty, we too can feel uncertain about which role or roles to adopt. Evaluation contributes to change and most would sign up to wanting evaluation to be useful and effective in the process. But how best can we do this?

In 2010, Luo[25] wrote about the proper role of the evaluator. He carried out a comparative analysis of five theorists' positions; explored how value, methods, use and purposes result in different roles for; considered how differences affect evaluators' responsibilities and proposed a resolution of the evaluator's role and as well as its limitations. Among the propositions sit the methodologist, the holder of authority, the educator and the facilitator. Luo points to the strengths and limitations of each set of defining characteristics and concludes that there is no single 'proper' role. It also appears that the relationship between evaluators and those they work with was not at the core of role. Luo's resolution argues that evaluators need to take account of different values and views; advise and agree what is acceptable design and methodology; ensure use of evaluation; and to acknowledge that making judgements is a central purpose of evaluation. All good and important yet I am not convinced this describes the evaluator who must function out there and face day-to-day challenges when on the job. So might the 'proper role' encompass the following:

25. Luo, Heng (2010) The Role for an Evaluator: A Fundamental Issue for Evaluation of Education and Social Programs International Education Studies

So might the 'proper role' encompass the following:

P: precise? personable?

R: rigorous? realistic?

O: organised? open?

P: principled? people-oriented?

E: evidence-focused? evidence-aware?

R: reasoned? reliable?

And what might it mean to be improper? Perhaps I: for innovative and M: for maverick!

There is a need to keep defining how we work; our roles and maximising usefulness. If evaluators and evaluation contribute to change, and I believe that is what we do, do our roles need to change so that we are not just the conduit of information and evidence but the brokers of relationships? Here stakeholders may feel more certainty that they have understood and helped create the evaluation despite the uncertainty within which we must all operate.

Co-production has become the go-to model. My experiences of these (often but not always well-intentioned) collaborations have ranged from exciting and rigorous to tokenistic observance of a seemingly acceptable approach, for example commissioning co-produced evaluation which turns into as top-down as they come and then, as evaluator, carrying the can when there is no data! And this has only encouraged me to put my co-producer hat back on to consider how we can maximise the strengths and minimise the challenges of co-production in evaluation. There are three (at a minimum) types of collaboration including the risk of 'cover-up' where the approach is tokenistic and focused on particular agendas. 'Critical collaboration' offers the critical friend in a safe space with safe hands. I would argue the need for a 'trusting collaboration' - a strong, confident, open-minded approach that puts others (sectors, organisations) ahead of self-interest. This requires the capacity to be adaptive, flexible and patient (is there a special hat for this?!).

Returning to how we bring out the best in evaluation, it seems important to look for opportunities for co-production. This can recognise power relations; acknowledge pluralistic perspectives and that evaluators can have several roles which depend on relationships with stakeholders. Checking in with people, checking in with ourselves and our peers about what we are doing and how we are doing it will help. Including capacity building in co-produced evaluation can support maximising interest and investment as people are less distanced from the main event. In our current situation, it seems increasingly important that we (co)create safe places within uncertainty and still do good and 'proper' evaluation.

Designing and delivering disability-inclusive evaluations: learning from the experience of DfID

ALISON POLLARD, DEPARTMENT FOR INTERNATIONAL DEVELOPMENT, MARK CAREW, SENIOR RESEARCHER - DISABILITY DATA & INCLUSIVE POLICIES, INTERNATIONAL LEONARD CHESHIRE AND LORRAINE WAPLING, INTERNATIONAL DISABILITY AND DEVELOPMENT CONSULTANT

The UK's Department for International Development (DFID) is committed to disability inclusive development. We published a new Disability Inclusion Strategy in 2018 and supported Leonard Cheshire to develop the Disability Data Portal. Through this article we reflect on why it is important for evaluators to work in ways that are disability inclusive. We share lessons learned from DFID's Girls' Education Challenge Fund, a scoping study about disability inclusive evaluation processes and systems (Wapling et al, 2017) and our collective experience.

Box 1: Definitions

The United Nations Convention on the Rights of Persons with Disabilities (UNCRPD): Persons with disabilities include those who have long-term physical, mental, intellectual or sensory impairments which, in interaction with various barriers, may hinder their full and effective participation in society on an equal basis with others.

The UK Equality Act 2010: A long term physical or mental impairment that has a substantial and long-term negative effect on your ability to do normal daily activities.

Prevalence and disadvantage associated with disability

Fifteen percent of the world's population has a disability (WHO, 2015). Not only are disabilities common, they are diverse: they may be visible, invisible and onset can be at birth, during childhood, working age or old age. The prevalence of disability increasing with age (WHO, 2015). People with disabilities play important roles in every society and are not defined by their impairment(s).

Data shows that people with disabilities are more disadvantaged than their peers without disabilities in terms of access to education, healthcare, employment, income, social support and are more likely to experience multiple deprivations (Mitra et al., 2013). These inequalities are a result of barriers, rather than any limitations of people with disabilities. Barriers to full participation in society include: attitudinal barriers; environmental barriers; institutional barriers; and inaccurate concerns over costs and difficulty of disability inclusion.

Applying a human-rights approach to disability

An individual model approach (often referred to as medical or charity models) is often applied to definitions of disability and to the design of programmes that aim to include people with disabilities. Interventions that apply an individual model focus on the person's impairment as the problem and seek to change individuals to help them 'fit in' to society primarily through the provision of impairment-based medical services. This has led to the segregation of public services which have greatly excluded people with disabilities from mainstream provision. Moreover, decision-making is led by 'specialists' rather than being driven by people with disabilities themselves with disability-led organisations rarely consulted. Working within this model can result in evaluators ignoring people with disabilities, assuming that their needs, experiences and rights are 'not relevant' to their work and not evaluating barriers that exacerbate disabilities and inequalities.

The human rights approach to disability follows the United Nations Convention on the Rights of Persons with Disabilities (UNCRPD) understanding **(see box 1)**. This acknowledges that disability results from the interaction between a person's impairment and the attitudinal, environmental and institutional barriers created by the social environment, which excludes people with impairments from participating on an equal basis with others.

The human rights approach promotes the understanding that people with disabilities have the right to participate in all development as active members of communities but may require adaptations to ensure their accessibility and inclusion. The implications of this approach are that different social agents must take responsibility for understanding what barriers exist for people with disabilities and institute measures to mitigate against them (Albert, 2004).

Disability inclusive data collection and analysis

Taking a rights-based approach implies that there are processes in place which can measure inclusion and track benefits specifically in relation to people with disabilities. This can pose particular challenges in the context of monitoring and evaluation. With disability recognised as a factor in development outcomes it needs to feature in plans and accountability mechanisms – whether at the international, bilateral, governmental, or programmatic level. Just as gender, age and ethnicity can be factors in marginalisation and vulnerability so too is disability and as such it needs to feature in the analysis of development outcomes. For that to happen, disability disaggregated data and disability inclusive evaluations are required.

We are working with all our partners to collect and analyse disability disaggregated data using the Washington Group Question Sets (**www.washingtongroup-disability.com**), which apply a human-rights definition of disability. As not all people have the same understanding of what disability means and because stigma is often associated with disability (making it unreliable simply to ask if a person or family member is disabled using a yes/no question), data on disability is prone to being inaccurate and of poor quality. It is important that questions used to obtain disability data are appropriately designed and implemented. The Washington Group Questions have been specifically designed to overcome these biases and do not use the term 'disability' or any related negative phrases. Instead they include cognitively tested neutral language focused on six core functional domains: seeing, hearing, walking, cognition, self-care and communication. Rather than capturing an individual's medical details or diagnosis the questions assess an individual's level of functioning in the six domains, making it possible to collect disability prevalence data without having to use any contested terms.

During data analysis, the evaluator creates a measure of disability from the Washington Group Questions by categorising the data dichotomously using a cut-off. The standard cut-off recommended by the Washington Group for comparable disability data is a score of "a lot of difficulty" or "cannot do at all" on any one of the six domains. Using the Washington Group Questions, evaluators can also look at the data using different cut-offs. For example, narrowing the cut-off for disability to "cannot do at all" will identify only those who experience severe functional difficulty, while extending the cut-off to include "some difficulty" will identify those that experience milder levels of functional difficulty. Data can thus be disaggregated and compared on key indicators using these different cut-offs. This would not be possible with a binary measure of disability.

Learning from the Girls' Education Challenge

The Girls' Education Challenge (GEC) was launched by DFID in 2012 as a 12-year commitment to reach the most marginalised girls in the world. It is the largest global fund dedicated to girls' education. The UK is committed to ensuring over a million girls in some of the poorest countries, including girls who have disabilities or are at risk of being left behind, receive a quality education. The GEC is currently supporting 40 projects in 17 countries. Over seventy percent of projects are working in Fragile or Conflict Affected States and all projects are collecting data using the Washington Group Questions.

A significant barrier to progress on disability inclusive education is the relative invisibility of students with disabilities within the education systems of low- and middle-income countries. By requiring projects to use the Washington Group set of questions within baseline, midline and endline data collection activities it has been possible to establish a reliable and consistent measure of disability prevalence amongst project beneficiaries.

Through analysing data from GEC baselines, we have found that projects which do not start out intentionally including students with disabilities among their target beneficiaries nevertheless report inclusion rates of between 3-5%. We have learned that the impact of requiring disability disaggregated data from all projects from the outset makes a significant difference to increasing the visibility of disability as a potential factor in predicting educational outcomes. By engaging with the measurement of disability, projects and their evaluators became more conscious and reflective regarding how educational outcomes are associated with disability in the particular context where they are working. This in turn has led to programmes reflecting on if and how they are working in disability inclusive ways and adapting their activities to ensure they become more disability inclusive.

Analysis of project responses to baseline data shows a broad range of activities taking place. Many projects opted to improve their own knowledge of disability, undertake awareness training, meeting with more experienced organisations, and engaging with disabled people's organisations. Some conducted more detailed barrier analysis to determine the nature of the exclusion facing girls with disabilities and many projects started to look at incorporating disability inclusive teacher training sessions into their support programmes.

Through our work we have learned that asking the following questions helps prompt critical and reflective thinking about how evaluators and commissioners of evaluations can work in disability inclusive ways:

- How can we address social barriers, including attitudes, about disability within our teams?
- How are we applying a human rights approach through the evaluation design, activities and budgets?
- Who is given a voice in data collection, data analysis and presentation and who is invisible? What action do I need to take as a result?
- How we can work in partnership with Disabled People's Organisations (DPOs) and in accordance with the principle 'nothing about us without us'?

Albert, B. (2004) Briefing note: The social model of disability, human rights and development. London and Norwich, UK: Disability Knowledge and Research Programme.

Mitra, S. Posarac, A. Vick, B. (2013) Disability and Poverty in Developing Countries: A Multi-dimensional Study. World Development, 2013. Vol 41 Issue C 1-18

Wapling, L. Buchy, M. and Resch, E. (2017): Scoping study: Donor Support for Disability Inclusive Country-led Evaluation Systems and Processes: Executive Summary, OPM

World Health Organisation, (2015) World Report on Disability, WHO

Stakeholder Involvement and Evaluation Influence: Putting Evaluation Theory to Practice for the Evaluation of Widening Participation Interventions in UK Higher Education Institutions

CATHERINE KELLY, UNIVERSITY OF BRISTOL SCHOOL OF EDUCATION

Regardless of whether an evaluation is classified as independent or participatory, it is often expected that stakeholders will be involved in the process in some way (Christie, 2003). This is especially the case in higher education institutions, which are required by the Office for Students (OfS) to evaluate the impact of their widening participation (WP) interventions and often do so internally. In 2015 the OfS announced the funding of a collaborative WP initially known as the National Collaborative Outreach Programme (NCOP). 29 consortia covering England were funded with the goal to increase higher education participation rates of young people aged between 13 and 18 who, after accounting for GCSE attainment, live in geographical wards with lower than expected higher education participation rates. This short article describes, from the perspective of the evaluator, the evaluation process conducted in one consortium and the effects of stakeholder involvement on perceived evaluation influence.

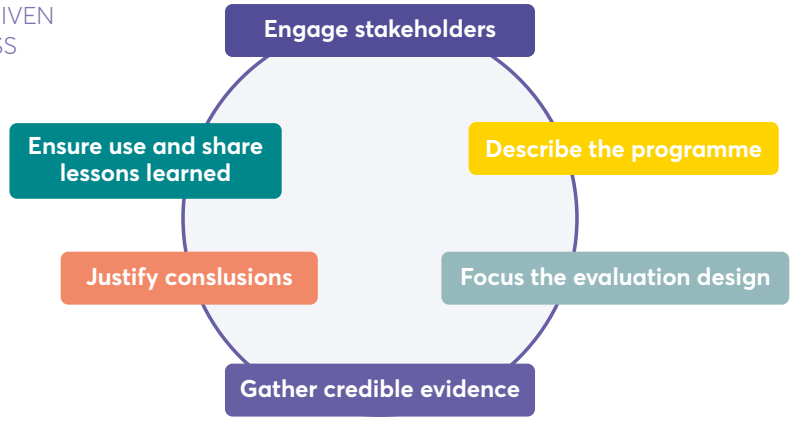
Background

Initially funded for 2 years, the consortium had approximately 25,000 target young people across 108 schools and 17 further education colleges. Collaboratively, with staff located across 5 universities and 17 FE colleges, 31 activities were developed to be delivered within the 2-year period. One local evaluator was appointed internal to the lead university and was required to develop and implement an evaluation strategy that could "demonstrate which interventions (in which contexts, and with which learners) have been instrumental in delivering progress, and which could have the most impact in the longer term" (HEFCE, 2017, p.10). The strategy needed to provide formative information for continuous programme improvement whilst supporting a longer-term summative research design that could feed into the national evaluation conducted by CFE Research.

Process

It was decided that in order to meet the summative and formative needs of the funders and programme staff, the best evaluation approach was theory-driven evaluation, centred on the development and testing of programme theories including theories of implementation, process, and outcomes. A theory driven approach follows three key processes, including identifying and engaging stakeholders to develop programme theory, formulating and prioritizing evaluation questions, and answering the evaluation questions (Donaldson, 2007). The consortium followed the evaluation steps presented in Figure 1. Key stakeholders included the consortium Steering Group and Operational Group, the Project Managers and Coordinators, OfS, CFE Research, the Higher Education Access Tracker who provided the database used to track and monitor students' educational outcomes, and the beneficiaries of the activities. Rather than the evaluator being in full control of how stakeholders were involved in the evaluation process, control over the evaluation strategy constantly swung back and forth between key stakeholders and the evaluator. Every decision implicitly or explicitly made, was the result of compromise and negotiation.

FIGURE 1: THEORY-DRIVEN EVALUATION PROCESS



Adapted from Centers for Disease Control Evaluation Framework (1999)

In the UK WP interventions are complex, they tend to consist of multiple components (Younger, Gascoine, Menzies & Torgerson, 2018), and in many cases, HE institutions are delivering similar interventions to the same students, making it difficult to isolate the effects of single interventions. Indeed, some activities delivered by the consortium were targeted, some were available to whole school year groups, some focussed on raising students' aspirations to progress to higher education, others on raising academic attainment and providing advice and guidance around routes into higher education courses. Furthermore, it is known that many other confounding factors influence young people's decisions about their education progression, particularly in their formative years in secondary school and college (Harrison et al., 2018).

In addition to the steps presented in Figure 1, the consortium had to:

1. Agree the legal basis for processing and storing personal student data required for tracking in the Higher Education Access Tracker.
2. Develop a monitoring system to track personal data of all participants required for the national evaluation of the NCOP.
3. Use programme theory to decide which programmes and activities to evaluate impact given resource and timing constraints.
4. Negotiate the methods for collecting data to measure prioritised outputs and outcomes without burdening beneficiaries.
5. Turn around formative findings in time for programme developers to improve programme implementation and outcomes.

Each of these decisions had implications for the quality of the evaluation and limited the methods we could use to rigorously answer prioritised evaluation questions.

Stakeholder Involvement and Evaluation Influence

Whilst the process of monitoring and developing a strategy for summative evaluation of resource intensive activities was challenging, often requiring the balancing of scientific rigour with feasibility, and negotiating and compromising to meet multiple and differing stakeholder needs, several key mechanisms of evaluation influence emerged. First of all, developing the programme theory highlighted the complexity of WP to programme stakeholders and the inherent complexities of implementing a research design that could rigorously test the impact of activities on students' progression to higher education. This process affected stakeholders' attitudes and beliefs of WP evaluation and opened space to discuss more innovative methods of data collection other than relying on subjective feedback ubiquitous in other WP evaluations (Harrison, 2012). The process also shed light on the importance of assessing implementation fidelity, particularly in the context of the consortium delivering activities across a large region and in a large number of schools and colleges. Furthermore, as a result of implementing a theory-driven, systematic approach to the evaluation, Programme Managers recognised the technical expertise required to meaningfully evaluate the programme and employed additional evaluation staff for the second phase of funding.

This example has shown that by virtue of following a systematic evaluation process and meaningfully involving stakeholders can affect changes in attitudes and beliefs of evaluation that can support the development of more scientifically rigorous and useful evaluations in the field of WP in UK higher education. **A key takeaway for WP** evaluators: persevere, respect stakeholders, and honour the integrity of the evaluation process (American Evaluation Association, 2018).

American Evaluation Association (2018). Guiding Principles for Evaluators. Retrieved from: https://www.eval.org/p/cm/ld/fid=51

Centers for Disease Control and Prevention. Framework for program evaluation in public health. MMWR 1999; 48 (No. RR-11)

Christie, C. A. (2003). What Guides Evaluation? A Study of How Evaluation Practice Maps onto Evaluation Theory. New Directions for Evaluation(97).

Donaldson, S. I. (2007). Program theory-driven evaluation science: Strategies and applications. New York, NY: Taylor and Francis Group.

Harrison, N. (2012). The mismeasure of participation: how choosing the 'wrong' statistic helped seal the fate of Aimhigher. Higher Education Review, 45(1), 30-61.

Harrison, N., Vigurs, K., Crockford, J., McCaig, C., Squire, R., & Clark, L. (2018). Evaluation of outreach interventions for under 16 year olds: Tools and guidance for higher education providers. Retrieved from Office for Students: https://www.officeforstudents.org.uk/media/e2c5eea5-b262-4ff6-8261-5b0bc84ba46a/ofs2018_apevaluation_a.pdf

Higher Education Funding Council for England HEFCE c (2017). National Collaborative Outreach Programme Guidance for Consortia (Rep.).

UK Evaluation Society

We're here to support the future of evaluators by promoting and improving the theory, practice, understanding and utilisation of evaluation

Discover more:
evaluation.org.uk
