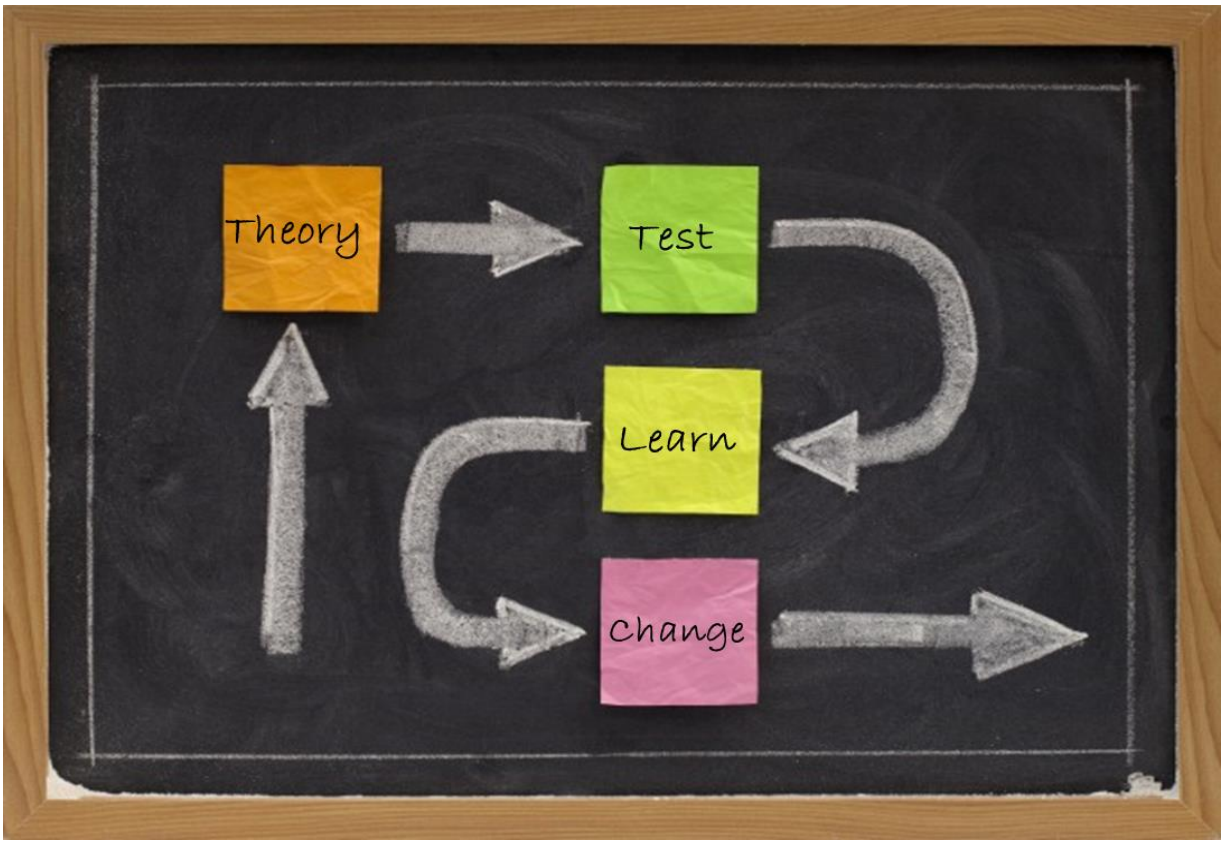
VIEWPOINTSERIES

JANUARY 2015

NEW YEAR'S RESOLUTIONS FOR THE INTERNATIONAL YEAR OF EVALUATION



JONATHAN COOK

SQW

FOREWORD

For a long time the Cinderella of the policy cycle, recent years have seen a rise in the prominence of evaluation in the UK. However, the increasing demands on learning 'what works' bring challenges on how to align policy-making and evaluation, and also debates over methods and practice. It is perhaps timely that this year is the International Year of Evaluation (EvalYear 2015), which aims to advocate and promote evaluation and evidence-based policy making at international, regional, national and local levels.

So, at the start of this special year for evaluation we share some of our own hopes for evaluation practice over the coming 12 months. We hope you will find our first Viewpoint of 2015 stimulating and thought-provoking and welcome any feedback that you may have.

Chris Green
Chief Executive, SQW Group

cgreen@sqwgroup.com

THE VIEWPOINT SERIES

The Viewpoint series is a series of 'thought piece' publications produced by SQW and Oxford Innovation, the operating divisions of SQW Group.

The aim of the Viewpoint series is to share our thoughts on key topical issues in the arena of sustainable economic and social development, public policy, innovation and enterprise with our clients, partners and others with an interest in the particular subject area of each paper. In each Viewpoint, we will draw on our policy research and implementation experience to consider key topical issues, and provide suggestions for strategic and practical solutions.

INTRODUCTION

We are now just over 12 months on from when the National Audit Office (NAO) published its report on the state of evaluation in government[1], and alongside it a separate review of government evaluations conducted by the London School of Economics (LSE)[2]. The NAO report was particularly critical of the evaluations commissioned by some government departments, making a series of recommendations relating to:

- the coverage of government spend by evaluation
- making evaluations more robust in how they assess impact
- the dissemination and use of evaluation evidence
- the transparency and prioritisation of evaluations.

The NAO's report painted a sobering picture of evaluation coverage and quality, highlighting the limited numbers of reports providing cost-effectiveness data, and/or sufficient evidence on policy impact. This in itself was not surprising, and it echoed our own work reviewing evaluation evidence over the last decade. White *et al.* (2008)[3] drew on reviews of evaluation evidence in the field of business competitiveness to highlight that very few evaluation studies comprehensively covered the different components of additionality (i.e. the difference made by interventions), and that the majority of evaluations were reliant on individual memory and self-reported benefits (rather than using comparison or control groups of non-beneficiaries).

Encouragingly, before and since the NAO report, there has also been more attention given to evaluation practice, in particular to the use of empirical methods. For example: The Magenta Book, HM Treasury's guidance on policy evaluation, was refreshed and republished in 2011 (with SQW contributing to the editing process); government has established a network of 'What Works' centres to disseminate evidence and help drive up the quality of evaluations; and there has been more emphasis given to randomised controlled trials (RCTs), encouraged in particular by the Behavioural Insights team, which was set up by Cabinet Office and is now part-owned by Nesta. We have also seen government departments taking a closer look at evaluation methods, in particular in relation to how best to undertake robust assessments of impact.

At this juncture, and at the start of the International Year of Evaluation, which aims to advocate and promote evaluation and evidence-based policy making at international, regional, national and local levels, it is useful to take stock as to the prospects for evaluation for the next 12 months and beyond. Here we present four key imperatives that we think the evaluation and policy-making community need to consider and respond to.

---

[1] National Audit Office (2013) Evaluation in government, NAO: London
[2] Gibbons, S., McNally, S. and Overman, H. (2013) Review of Government Evaluations: A report for the NAO
[3] White, G., Cook, J., Dickinson, S. and Heuman, D. (2008) Making the Most of Evaluation, SQW Viewpoint Series

BROADENING WHAT IS MEANT BY THE 'GOLD STANDARD'

For some in the research community, RCTs are identified as the 'Gold Standard' in evaluation practice. The Maryland Scale provides a scoring method from 1-5 to inform how close to this Gold Standard an evaluation has reached, with 1 representing the use of simple cross-sectional correlations and 5 the use of an RCT. The NAO and What Works Centres have used the Maryland Scale to determine the usefulness of evaluation evidence in terms of assessing impact, and we are now seeing government departments using similar means for reviewing evaluations. Whilst this provides a relatively straightforward and transparent scoring of evaluation quality and is appropriate for a number of policies and their associated evaluations, it is not appropriate to evaluate all interventions using RCTs, or even by using empirical evaluation methods. Indeed, the Magenta Book itself highlights certain situations in which empirical impact evaluations (based on econometric or statistical analysis of treatment and non-treatment groups) may not be feasible. It suggests the following might be situations where empirical evaluation is not appropriate[4]:

- There is a complex relationship between the outcome of interest and the driver of interest, with many other influencing factors. Complex interventions may occur in many policy domains and will be of increasing relevance as we respond to multi-faceted societal challenges, such as climate change and an increasingly ageing population.
- There is a long time lag before effects become realised, which can make demonstrating cause and effect challenging (and may also be too late to be of use in informing policy-makers). For example, this may be particularly relevant in business R&D and innovation, given the time lags to commercialisation.
- Targeting of interventions is highly subjective. This is particularly the case where decisions on who benefits are on the basis of defined criteria, such as in response to competitions for funding, or because of fit with a particular scheme (e.g. only eligible for certain technologies/ideas).

It seems that if we are to score evaluation evidence using a Maryland Scale, and potentially, as a result, focus more on interventions that can be evaluated with a score of 4 or 5, we risk not investing in complex interventions or those that have an effect only in the long-term. This would be folly given the nature of the challenges facing policy-makers and society more broadly. Related to this, the current focus on RCTs needs some reconsideration. Yes, they are a vital part of the evaluation toolkit, and should be adopted where it is appropriate to do so, but they are just that, "part of the toolkit". The Innovation Growth Lab, a joint initiative between Nesta and Kauffman, focusses entirely on RCTs. Whilst it is encouraging for the evaluation and policy-making community to think about using RCTs, it represents the research methodology driving the intervention, which, we would argue, is inappropriate.

[4] *HM Treasury (2011) The Magenta Book Guidance for evaluation, London, p101*

At a recent event on RCTs, breakout groups discussed the potential for using RCTs to evaluate current/potential initiatives. Only one group came up with a substantive RCT idea; for the remainder it seemed that there were barriers to feasibility or there were more appropriate means of evaluating the initiatives under consideration.

Therefore, if there is a 'Gold Standard' for evaluation, it should be defined as the best method for evaluating the particular intervention of concern. It should be the problems/challenges we face that drive the intervention, and the intervention should then be designed in such a way as to develop the best evidence on the solution, whether this design be an RCT, a quasi-experiment or a theory-building study. Indeed, it could be that combinations of methods are appropriate. For example, theory-based approaches can usefully complement empirical methods, because they enable one to understand more about the reasons for effectiveness (or otherwise) of interventions, which is particularly important where behavioural change and psychological factors are relevant. They can also be set up to provide greater real time feedback, which can improve the relationship between evaluation and policy-making, an issue to which we now turn.

IMPROVING THE RELATIONSHIP BETWEEN POLICY-MAKING AND POLICY-LEARNING

Evaluation evidence is not used by policy-makers as much as it ought to be. There are key barriers to this, not least the past quality of evaluation evidence and the fact that evaluation evidence (from their point of view) comes too late to inform design effectively. From the perspective of evaluation analysts, the design of policies (including pilots) hinders the delivery of good quality evaluation.

Therefore, there exists something of a 'chicken and egg' situation. Policy-makers may not use evaluation evidence because it is perceived to be of insufficient quality, or arrives too late. However, this can only be addressed if the relationship between the policy-making and the policy-learning processes are improved. Policy needs to be designed in ways that mean evaluation can be implemented robustly. This might require, for example, a shift from the existing situation whereby some programmes are piloted and then rolled-out before any evaluation can realistically be undertaken to inform wider implementation. This would mean that learning from evaluation could be incorporated into interventions and help to avoid perpetuating programme errors. In order to do this, there needs to be buy-in from policy-makers and ministers to the results of evaluation, and drawing in evaluation thinking at the time that interventions are designed.

In addition, can evaluation be designed in ways to provide better real time feedback, in particular for some interventions where it will be a number of years before outcomes are realised? The political reality means that policy-makers need to make certain decisions whilst they wait for the evidence of an empirical study. This highlights a key role for methods that gather feedback on how interventions are doing (e.g. through process evaluation) and also on whether the outlook is good for the achievement of outcomes. The assessments of outcomes could use theory-based approaches that gather evidence on the individual links of the theory/logic in real time, and include greater specification of intermediate, as well as final, outcomes. Whilst imperfect on cause and effect, done well these can be used to assess the 'plausible contribution' of the intervention, and can indeed be used alongside empirical approaches.

IMPROVING TRANSPARENCY AND SHARING OF KNOWLEDGE

The NAO indicated the need for more transparency in the reporting of evaluation findings. However, there can be a reluctance to publish reports, because independent advisors have been critical of the absence of RCTs and the like, even though there are clear reasons why they have not been used. It seems perverse, but the desire by departments to be more challenging of their own evaluation work has led to an adverse effect in the policy-learning process, which is the ultimate aim of evaluation. This unintended adverse effect may have been picked up from a theory-based approach to the assessment of evaluation practice!

So, how can transparency be improved? Some commentators have suggested independent bodies be established to commission, deliver and/or advise on evaluation in different domains. There are some examples of this, e.g. the Education Endowment Foundation (EEF) and the National Institute for Health and Care Excellence (NICE). However, it is perhaps no coincidence that these represent policy domains where RCTs have been used more widely, and where RCTs can be more feasible. In other policy domains, interventions are often more complex and, with debates on methodology in full swing, such independent bodies may push us too far down inappropriate methodological routes. Indeed, this is still an issue in relation to EEF and NICE, with neither covering the full range of evaluations in their own policy domains, with some acknowledgement that their own focus on RCTs is not always appropriate.

More generally, it would be beneficial to government analysts, policy-makers and evaluation practitioners outside of government to share knowledge and practice. Some government departments have research and evaluation panels that they use for procurement purposes; these are currently underutilised for the purpose of sharing knowledge and disseminating challenges facing departments. Outside of government, networks and societies such as the UK Evaluation Society could also provide an effective forum for developing communities of practice.

RESPONDING TO THE DEVOLUTION AGENDA

The discussion so far is particularly relevant at national level, and in particular within central government departments and their agencies. The desire within English cities and counties for greater devolution of powers presents the fourth challenge for local level evaluation that we highlight in this paper.

At local level, local government and Local Enterprise Partnerships ought to take note of the issues being raised in evaluation practice. Whilst there are constraints on resources, there needs to be consideration of evaluation now, in terms of:

- what evaluation evidence can say about what works in delivering local programmes
- planning for evaluation of initiatives under, for example, Growth and City Deals and Strategic Economic Plans.

The newly-established network of What Works Centres, including one focussed on Local Growth, is being used as a source of intelligence for these issues. However, the econometric techniques that form the focus of these are unlikely to say that much on devolved models. On the second point, thinking and planning now may enable more fit-for-purpose monitoring and evaluation methods in the future that provide robust evidence on interventions. The motivation for doing this is the demonstration of value for money to both local and national policy-makers and funders, thereby justifying subsequent investment and resource.

Devolution also poses some challenges at national level. We have seen the use of section 31 agreements to fund national schemes that are delivered locally. This provides local partners, within certain parameters, some freedoms around how and when they use the funding. Similarly, Growth Deals provide local partners with greater flexibilities on the use of funding, and this will become more commonplace if there are greater devolved powers. For national government, however, these flexibilities can cause problems because initiatives are delivered in different ways in different places, data collection such as monitoring may also be undertaken in different ways, and initiatives may be altered to account for changing local circumstances. This is not wholly conducive to the design of empirically strong evaluation, but reflects the direction of travel under devolution. What are the solutions here? In addition to encouraging local partners to develop evaluation plans, three other points are pertinent:

- National government/agencies can provide advice/support up front to encourage monitoring, and specifically consistent monitoring between local areas.
- There can be a minimum requirement for some form of baseline to be established at the outset of programmes. This would at least provide evidence on the contextual conditions that could be compared against later.
- Those interventions where empirical evaluation is most feasible can be identified with appropriate evaluation resources dedicated to these. For those where this is less feasible, innovative use of econometric techniques may be possible in some cases, but in other cases counterfactuals of the kind advocated under the higher reaches of the Maryland Scale may not be possible. Therefore, other approaches may be required to provide evaluative feedback instead of, or alongside, empirical techniques.

SUMMARY OF 'RESOLUTIONS'

At the start of 2015, the International Year of Evaluation, we have identified four particular challenges for the evaluation and policy-making communities:

- There needs to be broadening of what is meant by the 'gold standard' to take account of circumstances where empirical approaches are not feasible or, indeed, desirable. This is important as policy-makers deal with complex challenges and consider interventions with long-term goals. As part of this, whilst policy design can be shaped by evaluation requirements, we should not use a specific evaluation methodology to drive our policy solutions.
- Policy-making and policy-learning need to work more closely together. On the one hand, policy-makers do need to be prepared to build in evaluation. On the other hand, evaluation needs to take account of the political reality and acknowledge the use of approaches that can provide evidence in real time of 'plausible contribution' before the results of empirical approaches are available.
- Transparency may be facilitated by independent bodies that can marshal the evidence. However, this needs to acknowledge the different standpoints on methods, and in particular challenges arising from evaluating complex interventions. Networks, societies and communities of practice can help to share and ensure debate on these standpoints.
- Looking ahead to a key political debate in 2015, devolution presents further challenges for the evaluation community to consider. Action to encourage consistent application of the basics such as monitoring and baselines will assist here. The complicated nature of initiatives under devolution is likely to require a suite of methods to be adopted.

About us

SQW and Oxford Innovation are part of SQW Group.
For more information: www.sqwgroup.com

SQW is a leading provider of research, analysis and advice on sustainable economic and social development for public, private and voluntary sector organisations across the UK and internationally. Core services include appraisal, economic impact assessment, and evaluation; demand assessment, feasibility and business planning; economic, social and environmental research and analysis; organisation and partnership development; policy development, strategy, and action planning.
For more information: www.sqw.co.uk

Oxford Innovation is a leading operator of business and innovation centres that provide office and laboratory space to companies throughout the UK. The company also provides innovation services to entrepreneurs, including business planning advice, coaching and mentoring. Oxford Innovation also manages three highly successful investment networks that link investors with entrepreneurs seeking funding from £20,000 to £2m.
For more information: www.oxin.co.uk

For more information about this Viewpoint, please contact

Jonathan Cook, Director, SQW
T:+ 44 (0)20 7391 4105 E: jcook@sqw.co.uk