ual: social design institute

Working Paper no. 3

Evaluation: Concepts and Practice

Patrycja Kaszynska
Senior Research Fellow, Social Design Institute (SDI)
University of the Arts London (UAL)

Contents

Abstract

This working paper looks at evaluation – as a documentation genre and a form of practice – as an object of inquiry with its own unique history and provenance, a specific structure and composition and a way of acting on the world. These are respectively addressed in the paper under the headings of genealogy, morphology and performativity. Firstly, in relation to genealogy the paper traces the material and discursive history of evaluation to its roots in the ideas of scientific control and predictability and, more shallowly, the pressures of policy making. Secondly, in relation to morphology it shows that evaluation is a construct and serves two — largely incompatible — goals: on the one hand it provides an instrument of commensurability, enabling comparison of different entities across different contexts; on the other, it provides a platform for case-specific exploration and in-depth learning. Thirdly, concerning performativity the paper discusses that evaluation has real socio-material effects which do not always - in fact rarely - overlap with those intended by the evaluators and the commissioners. The key message of this paper is that evaluation is an artefact constructed for the purposes of decision making. Rather than a representation of 'real' change from some neutral point of view, it is a tool to get things done. This curtails in some way the learning potential implicit in evaluation. This said, attending to the genealogy, morphology and performativity in evaluation paves a way for a more critical understanding of evaluation and more reflective use. This working paper concludes with stressing the importance of asking about the purposes of undertaking evaluation and factoring this into evaluation design. The importance of means-ends reasoning and the need for thinking conjointly about prioritising goals as well as measuring sizes are thus suggested as foundational for more reflective evaluation practice.

Introduction

Evaluation can be defined as the making of a judgement or calculation about the number, amount or value of some phenomena based upon accepted qualitative distinctions and measurements. Most typically, evaluation is concerned with capturing change — a 'difference' that the evaluated intervention makes.

This compelling simplicity, however, stops as soon as it is realised that there are different institutional discourses of change (the story can be told in different ways), the choice of measurement and assessment methods influences what is recorded, the way judgements are expressed is often linked to different agendas and the categories used to generalise may provide an uncomfortable fit for specific contexts. Indeed, what matters first and foremost is to realise that evaluation is constructed: it is not a representation of 'the real' change from some neutral point of view, using a universal toolbox. Evaluation as a form of assessment has its own discursive history that reveals it as an artefact shaped by different agendas, which has evolved in response to specific — mostly policy-dictated — needs.

As this discussion shows, evaluation is not just a complex but — to an extent — an internally conflicted artefact: on the one hand, deployed for the purposes of accountability and benchmarking; on the other, concerned with how programme-specific Theories of Change (ToC) play out in case-specific circumstances. In the first instance, it emphasises generalisability and measurement; in the latter, understanding and explanation. Evaluation can be used as a tool in administration and as an aid to learning. To complicate the situation further, evaluation has real socio-material effects that do not always overlap with those intended by the evaluators and the commissioners. These complexities are the topic of this paper. The argument is that, by becoming aware of these complexities, evaluation can be used more effectively. Accordingly, this paper considers where evaluation comes from (genealogy), what it is (morphology) and what it does (performativity), with a view to drawing practical conclusions for evaluation practice.

Genealogy

Evaluation has a 'shallow' and a 'deep' history. On the shallow historical reading, evaluation can be traced back to the rapid development of programme evaluation in the US in the late 1960s (Madaus, Stufflebeam, Scriven, 1983); the expansion of evaluations supported by the European Union (European Commission, 1999), and the arrival of New Public Management (NPM): a style of governance that has been responsible for introducing private sector management models into the public services in the UK from the 1980s. Under NPM, the reduction of state bureaucracy and the decentralisation of public agencies went hand in hand with results-based management. This, in turn, created demands for accountability and new ways of demonstrating impacts. As a result, the culture of data-collection and evaluation-production flourished (Selwood, 2002; Clements, 2007).

However, as pointed out by numerous commentators, "although the terms program evaluation and evaluation research are relatively recent inventions, the activities that we will consider under these rubrics are not. They can be traced to the very beginnings of modern science" (Rossi et al., 2018, p.13; see also Cronbach et al., 1980).

Like any historical object, evaluation has been shaped by different ideologies, intellectual 'schools' and positions (see below).

Intellectual 'schools' shaping the cultures of evaluation — key examples:

Positivism
Relevant claims: factual knowledge is gained through observation (the senses), measurement is an expression of real life properties, society and social relations can be measured objectively.

Instrumental rationality
Relevant claims: the application of reason and law is a way of safeguarding objectivity, the key task for organisations/institutions is to have well-functioning bureaucracies and data management systems.

Neoliberal ideology
Relevant claims: markets offer solutions to social issues that self-interested bureaucracies could not deliver, the management of the social can be done using the market and the attached delivery mechanisms, such as monitoring techniques.

The practice of evaluation in the UK reflects these theoretical developments: it speaks to the Positivist solidification of the status of a certain model of a scientific proof, while also responding to wider social changes, such as the development of Weberian administrations and management of the social. Crucially, while sharing the Weberian desire to rationalise administration and governance — rather than seeking to involve the state apparatus — it does so in a neoliberal way by assuming that social scientific techniques, such as monitoring and evaluation, can do the job of state bureaucracies in the 'old-style' government.

Morphology

Stufflebeam (2000) identifies 22 foundation models for programme evaluation. Indeed, taxonomies in evaluation are now a well-established area of scholarly inquiry and a topic of debate (Chen, 1996; Miller, 2010). Here is not the place to rehearse these discussions; it is useful, however, to revisit some basic distinctions insofar as understanding them allows us to see evaluation as a 'construct' underpinned by specific agendas and originating in specific contexts, rather than a simple technique for capturing the 'objective' effects of any given intervention.

Some basic distinctions

Many sources suggest that it is possible to speak of just two main approaches: formative and summative.

Formative evaluation has been defined as "an on-going process that allows for feedback to be implemented during a program cycle" (Boulmetis and Dutwin, 2014). Fundamentally, it is a type of evaluation conducted in parallel with the programme and intended to provide feedback to be fed back into the programming.

Common Types of Formative Evaluation

Needs assessment determines who needs the programme, how great the need is and what might work to meet the need.

Structured conceptualisation helps stakeholders define the programme, the target population and the possible outcomes.

Implementation evaluation monitors the fidelity of the programme delivery.

Process evaluation investigates the process of delivering the programme, including alternative delivery procedures.

Adapted from the Web Center for Social Research Method's "Research Methods Knowledge Base" (http://www.socialresearchmethods.net)

The second category, summative evaluation, is usually conducted at the end of a programme to provide an assessment of what has been achieved. As illustrated below, it comes in different varieties and, as will be explained further, can have different objects.

Common Types of Summative Evaluation

Goal-based evaluation determines if the intended goals of a programme were achieved, e.g. Has my programme accomplished its goals?

Outcome evaluation investigates whether the programme caused demonstrable effects on specifically defined target outcomes, e.g. What effect does programme participation have on students?

Impact evaluation is broader and assesses the overall or net effects — intended or unintended — of the programme, e.g. What impact does this programme, have on the larger organisation (e.g. high school or college), community or system?

Cost-effectiveness and cost-benefit analysis address questions of efficiency by standardising outcomes in terms of their monetary costs and values, e.g. How efficient is my programme with respect to cost?

Adapted from the Web Center for Social Research Method's "Research Methods Knowledge Base" (http://www.socialresearchmethods.net)

The basic formative-summative dichotomy is, however, difficult to maintain in practice and has been challenged, most notably by Chen (1996). To overcome some of the shortcomings of the dichotomy, Chen offers his own conceptual framework which, he claims, allows more complete classification of evaluation types. The proposed typology (reproduced below), is achieved by crossing two evaluation functions ('improvement' and 'assessment') with two programme features ('process' and 'outcome'). The four resulting types are claimed by Chen to be able to encompass the entire spectrum of evaluation practice.

Evaluation Functions

	Improvement	Assessment
Process	Process-Improvement Evaluation	Process-Assessment Evaluation
Outcome	Outcome-Improvement Evaluation	Outcome-Assessment Evaluation

← **Figure 1. Basic Types of Evaluation, source: Chen, 1996.**

What is significant about Chen's typology is that it shows that the basic 'construction' of evaluation requires a crossing over of two different axes; it makes explicit that what is at issue is a combination of objects and objectives. This can be illustrated well by looking at the following definition:

> Evaluation is the systematic assessment of the design, implementation or results of an initiative for the purposes of learning or decision making (Poth et al., 2014)

The definition highlights two axes: objects and objectives. The use of the disjunctive 'or' is also telling as it signals that, at least in practice, choices have to be made between a number of different objects (design, or implementation or results) and different objectives (on the spectrum of learning to decision-making). Different typologies have been proposed in recent scholarship — see for instance Schoenefeld and Jordan (2017) — however, the distinction highlighted by Chen remains well-accepted.

Objects and objectives

It is possible for a single evaluation to look at different aspects of any given programme, e.g. design, implementation, results, achievements, etc. Achievements in turn can be evaluated in terms of: progress against goals and expectations; effectiveness of the intervention in achieving those goals; quality of the achievements (Parsons, 2017). Thus, the objects of valuation (the things being evaluated) can take on many different forms. In principle, there is nothing to stop a single evaluation from considering a range of objects; however, in practice — and given resource limitations — it is rare for evaluations to look at a wide range of objects.

It is difficult to combine different objectives into a single evaluation — both in theory and in practice — because aiding decision-making on the one hand, and supporting learning on the other, reflect fundamentally different expectations about the purposes of the evaluation and require different design, tools and questions.

At the risk of over-generalising, evaluations geared towards decision-making tend to emphasise measurement, often in relation to set standards; their focus is on quantifying how much change occurred. While doing this, they often deprioritise understanding, which is concerned with interpreting how change came about and the reasons why an intervention made a difference (see Rossi et al., 2018, p.4; Scriven, 1991, p.139).

This fundamental difference in emphasis reflects different objectives, but it can be seen as symptomatic of more deeply entrenched binaries: interpretivism versus positivism, measurement versus worth, value versus impact. These in turn, as discussed below, reveal not just different epistemic positions but also different attitudes to how power relations are played out in and through evaluation practices. What is at issue are the claims of evaluators to know, to judge and ultimately to control, and the role of the dominant value regimes in dictating the rules of engagement.

Performativity

'Performativity' can be taken to mean that the introduction of concepts and categories shapes the reality in which they are used (see Austin, 1975; Butler, 1997; MacKenzie, 2007).

Evaluation, as a general category, is performative. Even before we speak of findings, ideas or generalisations intended to shape future action, the fact that evaluation is conducted can be seen performatively. Just as the sheer "availability of metrics for generating and ordering information hierarchically about performance creates a demand for such information" (Rijcke, et al., 2016, p.164), so too the very existence of evaluation creates its own demand.

It also shapes society in various ways. There is already a wealth of research on the dependence of modern governance on standards, statistics and benchmarking (Desrosieres, 1998; Thévenot, 2011), and how the configurations of these "complex managerial" institutions shape the social giving rise to the various forms of "audit society" (Power, 1997) and the "black box society" (Pasquale, 2015). In this context, the caricatures of "management by targets", "box ticking" and "governing through numbers" (Desrosieres, 1998) hint at the irony of delivering "measurable outputs that functioned well within mechanisms of oversight, but had problems in creating the outcomes that policy-makers desired" (O'Brien, 2015, p.84). In this way, evaluation can be seen to shape realities in much the same way as do policies which, as Bacchi (2009) suggests, "give shape to 'problems', they do not address them" (p.X).

Evaluation can also be said to shape subjectivities. As audit culture becomes inherent in nearly all aspects of human activity, the way one relates to oneself also changes. In other words, "an estimation of one's worth by one's self-reflexively using evaluation devices for this purpose" (Forseth et al., 2019, p.32) becomes a common practice. This solidifies the role that evaluation plays in relation to what Foucault dubbed "governmentality": a form of governance applying quasi-market mechanisms and evaluating devices to both inter-subjective and intra-subjective relations. Just like human interactions become self-regulating, so too are internal psychological processes subject to self-correction with the aim of producing hegemonic submission (Foucault, 2008).

The use of evaluation

This opens the debate about who, if anyone, controls what evaluation does. Notably, two important evaluation scholars have shaped this discussion: Patton and Weiss. Patton emphasises a rational model of decision-making focusing on individual

decision-makers and the idealised notion of 'utilisation' by which he means "intended use by intended users" (Patton 2008, 2010). Weiss challenges the rational model of decision-making in favour of "a complex understanding of decision-making and policy-making in which evaluation findings are internalised, selectively used and rarely lead directly to specific decisions or changes in policy" (Descy, et al., 2004, p.34). Weiss argues that, because the phenomenon of use is highly complicated, it is more accurate to speak of "creep and accretion" rather than knowledge utilisation in decision-making (1980).[1] This is a reflection on the nature of policy-making rather than a specific observation about evaluation; still, it supports the point that evaluators have limited control over what happens to their 'findings'. As already pointed out, in relation to governmentality, the mere fact that evaluation is produced is not without consequences. To illustrate the same point differently, it could be pointed out that evaluation is a type of signalling giving legitimacy to programmes. As Weiss argues, evaluation can influence the "image of the program, even before any data or findings become available" (1980, p.25). Needless to say, the choice of measures, methods and design can be equally significant.

> The choice of measures (or in qualitative evaluations, the focus of study) can influence program operations. That is a way that evaluation is used. And not only the measures but also the design of evaluation itself can be "used". It may sound far-fetched that the very methods of study can have influence on subsequent program and policy events, but Breslau (in press) has written a fascinating account of how this has happened in the federal employment and training field. He shows how the use of evaluation was based, not only on the transmission of results, but on the evaluators' categories of data, design and analysis. (Weiss, 1998, p.26)

This brings us back to the morphology of evaluation: Is there anything inherent in how evaluation is constructed that makes it complicit with the regimes of governmentality and audit culture?

Balancing on the 'conflicted spectrum'

The contention of this paper is that evaluation can be thought of as a conflicted spectrum. Evaluation is, and has been used as, an instrument of accountability, advocacy and governmentality. It can be used, however, to further learning (Torres and Preskill, 1999) and multi-stakeholder engagement (Fazey et al., 2014).[2] These are not easy to reconcile and the problem is exacerbated by the fact that they are correlated with methodological and philosophical positions that can be constructed as opposing binaries (see below).

Evaluations prioritising measurement, generalisability and causal explanation
Focus is on impact measurement (what the intervention has typically produced, achievement vis-a-vis performance targets); assessment is criteria-based (concerned with the elucidation of regularities) and the criteria of value are externally imposed.

Prioritising explanation, understanding and hermeneutics
Key concern is not with quantifying how much change has occurred but with interpreting how change came about or the reasons why an intervention made a difference); assessment is instance-based (concerned with the observation of regularities) and the criteria of value are internally dictated.

Adapted from Meyrick, 2016 and Parsons, 2017

Throughout, this working paper has stressed the tension that "while information reduction affords cognitive efficiency and streamlines decision-making, its glossing of complexity, multiplicity and ambiguity can generate unintended consequences" (Orlikowski, et al., p.872). The difficulty, therefore, is in finding a way to generalise across different contexts in a way that preserves enough context specificity not to distort the object of evaluation and to keep track of the crucial differences.

There are some evaluative approaches relevant to this predicament: notably, "realist evaluation" developed by Pawson and Tilley[3] and "outcome mapping" popularised in the 2001 book by Earl, Carden and Smutylo but practiced in different variations as a less theoretically sophisticated and descriptively focused answer to the problem of preserving context specificity.[4] As Appendix 1 shows in relation to ToC — which for present purposes can be treated as an analogy for evaluation — there are different ways of responding to this challenge.

These efforts notwithstanding, reconciling the need for generalisability across different contexts with the need to account for the context specificity of individual interventions remains methodologically and conceptually challenging. It is also a politically sensitive issue. As the generalising functions of evaluation become dominant, the aspects of evaluation which enable it to support learning and participatory processes (Silvani et al., 2005; Molas-Gallart, 2012) become diminished. Conflicts, antagonisms, tensions and ambivalence — Mouffe's essential "impurities" of democracy — vanish as the pursuit of accountability accelerates.

At this point it would be easy to endorse one end of the spectrum and to dismiss the demand that evaluation is used for the purposes of accountability and governmentality. However, reneging on the accountability end of the spectrum would be to forget the purpose of evaluation: unlike the more open-ended forms of social science research[5], evaluation is supposed to facilitate decision-making. In the words of Espeland (2001), "we cannot make decisions unless we take seriously our cognitive limitations and our need to simplify our decisions" (p.1844).[6] Evaluation is the device invented to serve this very purpose. So, the question is not whether but how to achieve balance on the spectrum from accountability to understanding.

One way of answering this question is to look at how evaluation can communicate with policy in terms that can be accepted at both ends of the spectrum: efficient decision-making across different contexts and enhanced understanding of

individual situations.[7] (The potential and limitations of policy-making vocabulary is addressed in a separate working paper in this series). Another approach is to ask how mainstream evaluation practice can be made more reflective and reflexive.

Reflexivity in evaluation

In the words of Margaret Archer, reflexivity is defined as the "regular exercise of the mental ability, shared by all normal people, to consider themselves in relation to their (social) contexts and vice versa" (Archer, 2009, p.1). The argument of this paper is that an understanding of the genealogy and morphology of evaluation and being aware of its performative effects can help with using evaluation more reflexively. It is also likely to make us more reflective evaluators. Key points are:

Evaluating has effects but not necessarily those intended by evaluators

Far from being an innocuous technique with no effects on the real world, evaluation is a powerful instrument that influences external affairs. Evaluation is performative and the mere fact that it exists is significant in terms of influencing social behaviours. Thus, evaluating has effects, although — as discussed above — not necessarily those intended by the evaluators. As Weiss (1980) points out, it is more accurate to speak of "creep and accretion" rather than knowledge utilisation in policy decision-making, as far as evaluation is concerned. This situation is unlikely to change and hence, as responsible evaluators, it is important to understand that, in undertaking evaluations, we are tapping into and contributing to a much wider knowledge utilisation culture which does not necessarily follow the model of rational decision-making.

Evaluation is undertaken for different reasons and with different objectives in mind

Distinguishing between different purposes in evaluation is a very productive first step. A number of scholars have analysed the discourse with this in view, including the evaluators at the Tavistock Institute (Stern, 2004; Descy and Tessaring, 2004), and observed that there are at least four distinctive objectives for doing an evaluation: (a) accountability, where the intention is to give an account to sponsors and policy-makers of the achievements of a programme or policy; (b) development, where the intention is to improve the delivery or management of a programme during its term; (c) knowledge production, where the intention is to develop new knowledge and understanding; (d) social improvement, where the intention is to improve the situation of the presumed beneficiaries of public interventions. These have been elaborated usefully by Descy and Tessaring in the table reproduced below on p.12 (Figure 2). Being transparent and explicit about what one wants to achieve is a good step forward to better evaluation practice.

Not all 'theories' are equal

There has been a strong bias within evaluation to focus on method, technique and — to a lesser extent — methodology. There is also, however, a long-established interest in the role of theory in evaluation (Chen, 1990; Shadish et al., 1991). This literature recognises that theory refers to different constructs (see below). In practice, however, the concept of 'theory' has been used indiscriminately to mean different things. The most fundamental distinction is that between explanatory

Purpose	Stakeholder	Focus	Main evaluation approaches	Key questions
Accountability for policy-makers	Parliaments, Ministers, funders/sponsors, Management Boards	Identifying constraints. How they should be overcome? Delivery and implementation strategies	Relating inputs to outputs, qualitative description, following processes over time	How well is the programme being managed? Can it be implemented better?
Development for programme improvement	Project coordinators, partner organisations, programme managers	Identifying constraints. How they should be overcome? Delivery and implementation strategies	Relating inputs to outputs, qualitative description, following processes over time	How well is the programme being managed? Can it be implemented better?
Knowledge production and explanation	Programme planners, policy-makers, academics	Dissemination of good practice. What works? Organisational change	Experimental and quasi-experimental studies, case studies, systematic reviews and synthesis	What is being learnt? Are there any lessons that can be applied elsewhere? How would we do it next time?
Social improvement and social change	Programme beneficiaries and civil society	To ensure full involvement, influence and control by citizens and affected group	Stakeholder involvement, participative advocacy	What is the best way to involve affected groups? How can equal opportunities and soical inclusion be ensured?

↑ **Figure 2. Types of evaluation, source: Descy and Tessaring, 2004.**

theories that attempt to provide generalisable and verifiable knowledge about the world on the one hand, and programme theories which guide specific interventions (Donaldson and Lipsey, 2006). As explained in 'The specific case of Theory of Change' (see Appendix 1), the generality presupposed in the 'theory' in the first sense can be the opposite of the programme specificity in the latter. Observing these distinctions is crucial.

Five bodies of theory relevant to evaluators

(a) Theories of evaluation: these would include programme theory; theories of change approaches and realist approaches which emphasise the identification of mechanisms underlying successful change which have to be understood in specific contexts and settings.

(b) Theories about evaluation: there is a growing literature on evaluation practice, use, design and capacity. Included in this category would be particular aspects of practice identified by Shadish et al. such as theories of valuing.

(c) Theories of knowledge: including the main debates about the nature of knowledge, epistemology, methodology, etc., and about the nature of causal inference.

(d) Domain and thematic theories: which could be described as a theory of the evaluation object. This would include bodies of theory about domains such as human resource development, skill acquisition, the development of human capital and equal opportunities that could inform evaluation design, programme/policy implementation and outcomes.

(e) Theories of implementation and change: often seen as relevant by evaluators, including understandings of policy change, the diffusion of innovation and administrative behaviour. Such bodies of theory are likely to condition the success of programme interventions and can be quite separate from the kind of programme theories referred to above.

From Descy and Tessaring, 2004, pp.32–33.

What is 'enough but not too much' context specificity?

Context-specificity is generally a problem from the point of view of decision-making. A particular example from policy would be cost-benefit analysis (CBA) which is a technique that requires the costs and benefits associated with any given policy to be compared with each other using a common standard or metric, i.e. money. That said, comparing between different contexts outside of a clinical environment has been shown to be notoriously difficult (Deaton and Cartwright, 2018). While multi-dimensional indices and balanced scorecard approaches can be seen as a way of addressing this challenge, they do not presently solve the problem (Markusen, 2013). Reflexive and reflective evaluation practice means being aware that, in most circumstances, the more generalisable the evaluation findings are, the greater the risk of losing validity (ensuring that what is purported to be measured is actually the thing being measured). As hinted here, evaluation approaches such as "realist evaluation" and "outcome mapping" start to address these problems and the use of some form of outcome tracking, combined with other evaluative approaches, is advisable.

Evaluation is political

There is an image of evaluation rooted in Positivism (Hawkesworth, 1988; Bamberger, et al., 2006) which holds that "evaluation produces factual data about societal structures and processes by employing concepts and methods borrowed from the natural and physical sciences […] Because such information transcends historical and cultural experiences, it is assumed to have political and moral neutrality" (Bovens, Hart and Kuipers, 2006, p.325). That such objective understanding is possible has been questioned, however, by key theories associated with social constructivism, notably Guba and Lincoln (1989), and Fischer and Forester (1987). These researchers argue that meanings are the primary objects of evaluation and not facts. In other words, what is evaluated is an interpretation of something that happened and not what happened as such. In this sense, rather than mirroring 'facts', evaluation reflects agendas, points of view and vested interests. To admit this is to recognise that evaluation is political — in the words of Bovens, Hart and Kuipers (2006) — "at the stage of agenda setting, problem definition and the selection of instruments some groups, interests and voices are organised 'in' the design and management of evaluation proceedings, whereas other stakeholders are organised 'out'" (p.325).

Concluding reflections

As this discussion reveals, evaluation is not just a complex construct but also, to an extent, one that is internally conflicted. It can serve two — largely orthogonal — goals: on the one hand it provides an instrument of commensurability, enabling comparison of different entities across different contexts, and hence supporting the objectives of decision-making; on the other, it can also provide a platform for in-depth exploration and case-specific learning. While these goals should not be assumed incompatible in principle, in practice, satisfying them pulls in opposite directions. The big challenge for evaluation practice is working out how to aid the process of policy-making without destroying or distorting the contextual information in the process.

The use of evaluation is equally complicated: it can be deployed strategically but it is rarely utilised in line with the intentions of the evaluators. Evaluation is not a representation of 'real' change from some neutral point of view. As a political and powerfully performative construct, it should be practiced with care. Far from being a mere, value-neutral tool, evaluation is a powerful instrument of control embodying different agendas. This does not mean that we should stop evaluating. Indeed, as illustrated, some form of assessment is essential to get things done. However, a clear understanding of the morphology and performativity of evaluation can help us to become better evaluators.

Accordingly, rather than treating evaluation in essentialist terms as a representation of an objective reality beyond it, this paper considers evaluation to be a device to get things done, "in action" as Latour would say (Latour, 1987).

Appendix 1. The specific case of Theory of Change

Theory of Change (ToC) is a tool to make explicit the 'logic' of how an intervention is expected to produce results. Davies defines ToC as "the description of a sequence of events that is expected to lead to a particular desired outcome" (Davies, 2018); Dhillon and Vaca as "the hypothesis about the way that a program brings about its effects…essentially the logic behind an intervention" (Dhillon and Vaca, 2018). Just like evaluation, ToC can be used at different stages and with different scopes (Rehfuess et al., 2018). Interestingly, there is also some acknowledgement that ToC can be used with different intentions, notably, for the purposes of partnership, team-building and communication (see below).

Purposes of ToC

Strategy: help teams work together to achieve a shared understanding of a project and its aims; make projects more effective; help to identify and open up 'black boxes' in thinking.

Measurement: help determine what needs to be measured (and what does not) in order to plan evaluation activities; encourage teams to engage with the existing evidence base; act as the basis for claims about attribution.

Communication: quickly communicate a project's aims; bring the process of change to the forefront.

Partnerships: help with partnership working.

Adapted from Harries et al. (2014) following Stein and Valters (2012)

Yet, in practice, these distinctions — in particular where objectives are concerned — tend to be omitted (Dhillon and Vaca, 2018; Davies, 2018). This commonly leads to the assumption that ToC is used as an 'image' of the proposed intervention and its effects, with the links in the chain underpinning ToC constituting causal pathways and the 'theory' in question having scientifically explanatory potential (Coryn, 2011; Donaldson and Lipsey, 2006). This is problematic. To use Mulgan's pithy formulation, a theory "is generally taken to mean an idea, principle or law that is separate from, and more general than, the thing being explained [but] the advocates of theories of change use the word 'theory' in an almost opposite sense to describe a specific explanation of a specific example (albeit one that should then have predictive power), but never explain why they do so".[8] Furthermore, if the de facto way in which ToC is used is to guide action, rather than to theorise the presumed changes, then 'for' would be more appropriate than 'of'. Finally, when it comes to the 'change' that ToC captures, it is not clear that most interventions can isolate the consequential changes attributable to the interventions.[9]

Very pertinently, in relation to reflexivity, questions have been raised about ToC's propensity to 'squeeze' both politics and learning out of evaluation practices. The former because of the implicit bias towards consensus and thus the pressure to produce one model of ToC that is consensually embraced by those involved in

the programme delivery (Leeuw and Donaldson, 2015).[10] With regard to the latter, it is because having a mapped-out ToC can pre-empt further reflection (Bensimon, 2012). Like design thinking, ToC has become a victim of its own implementation success: it is an example of a tool used on mass scale, but mostly in a non-critical way. These shortcomings are, however, not inevitable.

Indeed, there is a growing body of discourse showing that ToC can be used to encourage creativity (Stein and Valters, 2012) and operationalised as a means of "structured experiential learning" by seeking to build "learning objectives into the cycle of project design, implementation, completion, and evaluation" (Pritchett et al., 2013).

How ToC are constructed and used in practice depends on the context of implementation:

> It is clear that the way in which Theories of Change are approached is closely related to the prevailing development discourses of 'results' and 'evidence'. With this comes a considerable danger that the approach will privilege a linear cause and effect narrative of change. There appear to be two schools of thought on the direction of Theories of Change: one which seeks to use the tool to expand our understanding of change contexts, and another which views them as a "logframe on steroids".[11]

Just like evaluation in general, ToC specifically should be subject to reflexive practice.

Endnotes

[1] This has been supported by Landry et al. who find that the determinants of knowledge utilisation lie in the linkages between researchers and users (what they label the "interaction model") instead of the nature of the research results ("science push"), the users' needs and context ("demand pull") or the efforts made to disseminate results ("dissemination model"). Thus, they agree with Weiss that knowledge utilisation depends on various disorderly interactions (Landry et al., 2001).

[2] In this context it is useful to draw attention to unified versus theory-based stakeholder evaluation (TSE). Instead of fusing into a single unified narrative, the latter keeps the ToCs of diverse stakeholder groups separate from each other and from the programme theory embedded in the institutionalised intervention itself.

[3] As Pawson and Tilley summarise the logic of realist explanation "The basic task of social inquiry is to explain interesting, puzzling, socially significant regularities (R). Explanation takes the form of positing some underlying mechanism (M) which generates the regularity and thus consists of propositions about how the interplay between structure and agency has constituted the regularity. Within realist investigation there is also investigation of how the workings of such mechanisms are contingent and conditional, and thus only fired in particular local, historical or institutional contexts (C)" (Pawson and Tilley, 1997; p. 71).

[4] The 'originality' of this approach lies in its shift away from assessing the products of a programme articulated in terms of external indicators, to focus on changes in behaviour, relationships, actions and activities in the people, groups and organisations affected directly by the programme.

[5] As Helen L. Chen explains in her entry on Evaluation versus Research in the *Sage Encyclopaedia of Educational Research:* although evaluation is "a type of applied research that employs similar methodologies", it does so for a "different purpose, focus and audience" when compared to research (Frey, 2018, p.629). The fundamental difference lies in the purpose and motivation, as indeed is emphasised in this paper.

[6] This applies as much to evaluation and valuation. As Orlikowski and Scott (2014) point out, "Prior literature highlights processes of reduction, authority, asymmetry and reactivity associated with valuation" (p.27).

[7] From the point of view of the author of this paper, the question is whether comparability can be achieved without necessarily assuming commensurability (the process of comparing different entities in terms of a common metric), permitting "scrutiny of complex or disparate phenomena in ways that enable judgment" (Espeland and Stevens, 2008, p.415).

[8] https://www.nesta.org.uk/blog/whats-wrong-with-theories-of-change/

[9] In this sense, most interventions struggle to account for unintended consequences: displacement (positive outcomes as offset by a negative outcome of the same intervention); substitution (benefits for some groups happen at the expense of other groups; leakage (benefits that occur outside of the population group); deadweight (outcomes that would happen anyway); additionality (the net changes, outcomes or impacts over and above what is expected).

[10] In this context see unified versus theory-based stakeholder evaluation (TSE). Instead of fusing the theories of different stakeholders into a single unified theory, the latter keeps the programme theories of diverse stakeholder groups separate from each other and from the programme theory embedded in the institutionalised intervention itself.

[11] https://oxfamblogs.org/fp2p/theories-of-change-logframes-on-steroids-a-discussion-with-dfid/

References

Archer, M. S. (Ed.). (2009). *Conversations about reflexivity*. Routledge.

Austin, J. L. (1975). *How to do things with words* (Vol. 88). Oxford University Press.

Bacchi, C. (2009). *Analysing policy*. Pearson Higher Education AU.

Bamberger, M., Rugh, J., & Mabry, L. (2006). *Real World Evaluation: Working under Budget, Time, Data, and Political Constraints.* Sage.

Bensimon, E. M. (2012). The equity scorecard: Theory of change. *Confronting equity issues on campus: Implementing the equity scorecard in theory and practice*, 17–44.

Boulmetis, J., & Dutwin, P. (2014). *The ABCs of evaluation: Timeless techniques for program and project managers* (Vol. 56). John Wiley & Sons.

Bovens, M., Hart, P. T., & Kuipers, S. (2006). The politics of policy evaluation. *The Oxford Handbook of Public Policy*.

Butler, J., & Butler, K. C. (1997). *Excitable speech: A politics of the performative*. Psychology Press.

Chen, H. T. (1990). *Theory-driven evaluations*. Sage.

Chen, H. T. (1996). A comprehensive typology for program evaluation. *Evaluation practice*, *17*(2), 121–130.

Clements, P. (2007). The evaluation of community arts projects and the problems with social impact methodology. *International Journal of Art & Design Education, 26*(3), 325–335.

Callon, M., Rip, A., & Law, J. (Eds.). (1986). *Mapping the dynamics of science and technology: Sociology of science in the real world*. Springer.

Coryn, C. L., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation*, *32*(2), 199–226.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., ... & Weiner, S. S. (1980). *Toward reform of program evaluation* 3. Jossey-Bass.

Davies R. (2018). Representing theories of change: technical challenges with evaluation consequences. *Journal of Development Effectiveness 10*(4), 438–461.

Deaton, A., & Cartwright, N. (2018). Reflections on Randomized Control Trials. *Social science and medicine*, *210*, 86–90.

Descy, P., & Tessaring, M. (Eds.). (2004). *The Foundations of Evaluation and Impact Research: The Third Report on Vocational Training Research in Europe: Background Report* (Vol. 58). European Communities.

Desrosières, A. (1998). *The politics of large numbers: A history of statistical reasoning*. Harvard University Press.

Dhillon, L., & Vaca, S. (2018). Refining theories of change. *Evaluation*, *14*(30).

References

Donaldson, S. I., & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: Developing practical knowledge. *The handbook of evaluation: Policies, programs, and practices*, 56–75.

Earl, S., Carden, F., & Smutylo, T. (2001). *Outcome mapping: Building learning and reflection into development programs*. IDRC.

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie*, *49*(3), 401–436.

Espeland, W. N. (2001). Value-matters. *Economic and Political Weekly*, 1839–1845.

European Commission. (1999). Evaluation design and management. *Means collection: evaluating socio-economic programmes*. Luxembourg: Office for Official Publications of the European Union, Vol. 1.

Fazey, I., Bunse, L., Msika, J., Pinke, M., Preedy, K., Evely, A. C., ... & Reed, M. S. (2014). Evaluating knowledge exchange in interdisciplinary and multi-stakeholder research. *Global Environmental Change*, *25*, 204–220.

Fischer, F., & Forester, J. (1987). *Confronting values in policy analysis: The politics of criteria*. Newbury Park.

Foucault, M., Davidson, A. I., & Burchell, G. (2008). *The birth of biopolitics: lectures at the Collège de France, 1978–1979*. Springer.

Forseth, U., Clegg, S., & Røyrvik, E. A. (2019). Reactivity and Resistance to Evaluation Devices. *Valuation Studies*, *6*(1), 31–61.

Frey, B. B. (Ed.). (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. Sage.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Sage.

Harries, E., Hodgson, L., & Noble, J. (2014). *Creating your theory of change: NPC's practical guide*. New Philanthropy Capital, https://www.thinknpc.org/wp-content/uploads/2018/07/Creating-your-theory-of-change1.pdf

Hawkesworth, M. E. (1988). *Theoretical issues in policy analysis*. SUNY Press.

Landry, R., Amara, N., & Lamari, M. (2001). Utilization of social science research knowledge in Canada. *Research policy*, *30*(2), 333–349.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Leeuw, F. L., & Donaldson, S. I. (2015). Theory in evaluation: Reducing confusion and encouraging debate. *Evaluation*, *21*(4), 467–480.

MacKenzie, D. A., Muniesa, F., & Siu, L. (Eds.). (2007). *Do economists make markets?: on the performativity of economics*. Princeton University Press.

Madaus, G. F., Stufflebeam, D., & Scriven, M. S. (1983). Programme evaluation. *Evaluation models*, 3–22. Springer. Dordrecht.

References

Markusen, A. (2013). Fuzzy concepts, proxy data: why indicators would not track creative placemaking success. *International Journal of Urban Sciences*, *17*(3), 291–303.

Meyrick, J. (2016). Telling the Story of Culture's Value: Ideal-Type Analysis and Integrated Reporting. *The Journal of Arts Management, Law, and Society*, *46*(4), 141–152.

Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation*, *31*(3), 390–399.

Molas-Gallart, J. (2012). Research governance and the role of evaluation: A comparative study. *American Journal of Evaluation*, *33*(4), 583–598.

O'Brien, D. (2015). Cultural value, measurement and policy-making. *Arts and humanities in higher education*, *14*(1), 79–94.

Orlikowski, W. J., & Scott, S. V. (2014). What happens when evaluation goes online? Exploring apparatuses of valuation in the travel sector. *Organization Science*, *25*(3), 868–891.

Parsons, D. (2017). *Demystifying evaluation: Practical approaches for researchers and users*. Policy Press.

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Patton, M. Q. (2008). *Utilization-focused evaluation*. Sage.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. Guilford Press.

Pawson, R., & Tilley, N. (1997). An introduction to scientific realist evaluation. *Evaluation for the 21st century: A handbook*, 405–418.

Poth, C., Lamarche, M. K., Yapp, A., Sulla, E., & Chisamore, C. (2014). Towards a Definition of Evaluation within the Canadian Context: Who Knew This Would Be So Difficult?. *Canadian Journal of Program Evaluation*, *29*(1).

Power, M. (1997). *The audit society: Rituals of verification*. OUP Oxford.

Preskill, H., & Torres, R. T. (1999). *Evaluative inquiry for learning in organizations*. Sage.

Pritchett, L., Samji, S., & Hammer, J. S. (2013). It's all about MeE: Using Structured Experiential Learning ('e') to crawl the design space. *Center for Global Development working paper*, 322.

Rijcke, S. D., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation*, *25*(2), 161–169.

Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2018). *Evaluation: A systematic approach*. Sage.

Rehfuess EA, Booth A, Brereton L, Burns J, Gerhardus A, Mozygemba K, Oortwijn W, Pfadenhauer LM, Tummers M, van der Wilt GJ and others. (2018) Towards a taxonomy of logic models in systematic reviews and health technology assessments: A priori, staged, and iterative approaches. Research Synthesis Methods 9(1):13–24.

References

Schoenefeld, J., & Jordan, A. (2017). Governing policy evaluation? Towards a new typology. *Evaluation, 23*(3), 274–293.

Scriven, M. (1991). *Evaluation thesaurus*. Sage.

Selwood, S. (2002). The politics of data collection: Gathering, analysing and using data about the subsidised cultural sector in England. *Cultural trends,* 12(47), 13–84.

Silvani, A., Sirilli, G., & Tuzi, F. (2005). R&D evaluation in Italy: More needs to be done. *Research Evaluation 14*(3): 207–215.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice.* Sage.

Stein, D., & Valters, C. (2012). *Understanding theory of change in international development*.

Stern, E. (2004). Philosophies and types of evaluation research. *Evaluation and impact of education and training: The value of learning. Third report on vocational training research in Europe: Synthesis report*. Luxembourg: Office for Official Publications of the European Communities.

Stufflebeam, D. L. (2000). Foundational models for 21st-century program evaluation. In *Evaluation models*, 33–83. Springer.

Thévenot, L. (2011). Power and Oppression from the Perspective of the Sociology of Engagements: A Comparison with Bourdieu's and Dewey's Critical Approaches to Practice Activities. *Irish Journal of Sociology, 19*(1), 35–67.

Torres, R. T., & Preskill, H. (2001). Evaluation and organizational learning: Past, present, and future. *American Journal of Evaluation, 22*(3), 387–395.

Weiss, C. H. (1980). Knowledge creep and decision accretion. *Knowledge, 1*(3), 381–404.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation?. *American Journal of Evaluation, 19*(1), 21–33.

Social Design Institute, UAL
socialdesigninstitute@arts.ac.uk

The Social Design Institute champions social and sustainable design at University of the Arts London. Its mission is to use research insights to inform how designers and organisations do designing, and how researchers understand design, to bring about positive and equitable social and environmental changes. The Institute achieves its mission through original research, translating research through knowledge exchange and informing teaching and learning.

Published by
University of the Arts London
272 High Holborn
London
WC1V 7EY

www.arts.ac.uk
June 2021

Design: www.timhutchdesign.com