Evaluating unintended consequences: new insights into solving practical, ethical, and political challenges of evaluation

Abstract:

Evaluating complex interventions and policies is challenging. This is particularly true for the identification of unintended consequences, whether negative or positive. This paper uses data from a workshop with policymakers and evaluators to explore the evaluation of unintended consequences. We identify three main challenges for policymakers and evaluators: being able to identify and evaluate unintended effects, to avoid creating unintended effects, and being able to explain these effects. We discuss practical, political and ethical issues for each of these challenges, and identify recommendations for evaluators who want to consider unintended consequences. Firstly, use a broader range of methods to explore how policies play out. Secondly, use theory to plan evaluations, and thirdly discuss both methods and theory with relevant stakeholders to make these as useful as possible. We offer novel insights into recent debates about theory-led and co-produced interventions and policies.

L'évaluation des interventions et politiques complexes est délicat, particulièrement en ce qui concerne l'identification des effets imprévus, qu'ils soient négatifs ou positifs. Cet article utilise des données provenant d'un colloque avec des décideurs politiques et les valuateurs sur l'évaluation des effets imprévus. Nous identifions trois défis principaux pour les décideurs politiques et les évaluateurs: identifier et évaluer les effets imprévus, les éviter, et les expliquer. Nous examinons les questions pratiques, politiques et éthiques pour chaque défi, et proposons des recommandations pour les évaluateurs qui souhaitent considérer les effets imprévus. Premièrement, utiliser une plus grande variété de méthodes pour examiner les effets des politiques; deuxièmement, utiliser la théorie pour planifier l'évaluation; et troisièmement, discuter les méthodes et la théorie avec les parties prenantes pour les rendre aussi utiles que possible. Nous offrons de nouvelles perspectives informant des débats récents sur les interventions et politiques coproduites fondées sur la théorie.

Keywords: Unintended consequences; evaluation theory; coproduction; politics of evaluation; evaluation methods; ethics

Background: Since its inception, evaluation science has identified the challenges of evaluating complex interventions and policies, particularly in complicated (multi-level and multi-site) and complex adaptive (emergent, non-linear) systems (Clark, 2013; Patton, 1994; Pawson and Tilley, 1997; Rogers, 2008; Sanderson, 2002). These include practical, political, and ethical concerns, in addition to a multitude of methodological problems (Byrne and Callaghan, 2013). Despite these challenges, most people agree evaluation must still be prioritised. In addition to providing evidence about whether a policy has had an impact – expected or otherwise – it can identify promising ideas and failing ones, and contribute to incremental social change (Sanderson, 2000; Weiss, 1993, 1998). Importantly, it can enable democratic conversations about policy choices, by clarifying the political trade-offs and their implications (Weiss, 1993).

As we know, the practice of evaluation is a political act, whereby only certain programmes and outcomes are evaluated; "conveying the message that other elements in the situation are either unimportant or that they are fixed and unchangeable" [pp.100 (Weiss, 1993). Evaluations can be used to indicate or suggest at weaknesses in programmes (so making them easier to challenge or shut down), close off debate or narrow policy options (Sanderson 200, Gray and Jenkins 1995. Negative findings can be buried by clients, arguing for more time, more money, better evaluation tools (Weiss, 1993). Evaluations are almost never from the point of view of participants, or those

affected, and "might well lead to very different recommendations from those developed by an agency-oriented evaluator or a program official" (Weiss, 1993: 102).

What would happen if evaluations did reflect users' perspectives? We argue that evaluation must address the role of complexity in understanding social problems, social change, and evaluation of both. We often single out the intervention as the key independent variable, but this is a very reductionist way of thinking about social changes (Warren, 1973) which ignores structural and institutional structures which generate and sustain the problems of the target group. How this broader point should be addressed in terms of evaluation methodology is a topic of ongoing debate (Bonell et al., 2012; Egan et al., 2009), but it seems clear that many widely used evaluation techniques are not adequate to the challenges of situating interventions and policies in their socio-political contexts. In particular, the use of methods such as randomised trials tends to focus evaluators' attention on strategies which are well-defined and portable, and on proximal and readily quantifiable outcomes, even though an increasing body of evidence suggests that piecemeal approaches to create social change can surely not be effective (e.g. Adab et al., 2018).

For Weiss, the answer was more integrated polices with "serious examination of the basic problem, how it is defined, what social phenomena nurture and sustain it, how it is related to other social conditions and social processes, and the total configuration of forces that have overwhelmed past program effects" (Weiss 1998:103). Of course, what is possible for evaluators in practice – given constraints of resource availability, political feasibility and so on – may fall well short of this ideal scenario. Seeing interventions as ongoing processes within complex systems may open up new possibilities for a deeper understanding of intervention, but it also raises challenging theoretical and practical problems as to which aspects of the total intervention-system nexus should be the main foci of evaluation. Commentators have called for less ambitious goals, more useful measures, consideration only of the most potent forms and elements of programmes (Bevan and Hood, 2006; Bjørnholt and Larsen, 2014; Pawson, 2008). However, this raises the question of what the aims of evaluation should be in any given context – questions which, of course, may be the subject of disagreement between the different actors involved in the evaluation process.

The question of how to prioritise the evaluation process may be particularly challenging when it includes unintended consequences (UCs). As used here, UCs include both adverse effects and positive 'spillover' effects on outcomes or populations not envisaged by the originators of the intervention or policy. UCs have been recognised as a pervasive structural feature of social action since at least Merton (1936),[1] and as an important dimension of the evaluation of interventions since at least Hirschman (1967). They are widely if sparsely identified in the evaluation and policy literature (Allen-Scott et al., 2014; Bamberger et al., 2016; Bonell et al., 2015; Loke et al., 2007; Lorenc and Oliver, 2013; Mittelmark, 2014), but there has been limited discussion of how policymakers and evaluators respond to evidence of UCs in practice. As ethical practitioners and researchers, it is incumbent upon us to ensure that we do no harm, either directly or through wasting resources on ineffective interventions (Hawe, 2015b; Moore and Evans, 2017). However, in many policy fields the identification and monitoring of adverse UCs has been at best *ad hoc*, and there is little clear guidance for evaluators or practitioners as to how to identify and respond to evidence that policies are doing harm. Understanding UCs may also be of value in refining our understanding of intervention mechanisms and their relation to the systemic context, and pinpointing inadequacies in mid-range theories of intervention – potentially mitigating the fact that,

[1] Strictly speaking Merton speaks of *unanticipated* consequences, which is arguably a rather different idea {De Zwart, F. (2015). Unintended but not unanticipated consequences. Theory and Society, 44(3), 283-297.} However, his paper has widely been taken as the starting point for reflection on unintended consequences.

as Pawson says, “we are still inclined to launch makeshift interventional rockets without a solid theoretical base in social and behavioural science” (Pawson, 2018). More broadly, a clearer body of evidence on UCs could contribute to improved policy design and implementation.

The purpose of this paper, therefore, is to move forward our understanding of unintended consequences and their evaluation, particularly: exploring the role that evaluation (of UCs) plays in the policy process; methodological challenges associated with evaluation of UCs; and ethical and normative issues associated with unintended consequences. We do this by presenting novel empirical data about policy actors’ own perceptions of the challenges and opportunities involved in evaluating with consideration for unintended consequences.

Workshop on unintended consequences of policies

This paper reports the findings of a workshop with senior policymakers and evaluators to explore the unintended consequences of public health and social policies. Workshop participants (n = 14) were split into two groups: Researchers and evaluators, and developers and implementers. We aimed to explore participants’ views on how evaluation should take account of UCs, what the challenges and constraints might be, and the ethics and logistics of evaluating unintended effects. We also wanted to explore scenarios where UCs were identified in evaluation, to think about what happened next, and whether the learning was captured and fed back into policy-making. Other sessions explored more general issues about why UCs arise and how adverse UCs can be avoided or mitigated; these findings have been reported elsewhere (under review), with this paper focusing on the evaluation messages. The write-up below contextualises our results by situating the workshop findings into current debates about evaluation methods and systems (see Table 1 for a summary). Where possible, key points raised by participants have been discussed in conjunction with debates in the literature. Similarly, where possible, cases or examples raised by participants have been referenced, for the convenience of the reader, but it should not be assumed that these were provided by participants unless otherwise indicated.

We recognise that this is a relatively small number of participants, and that we have referred to ‘researchers’ and ‘policymakers’ as generic categories, as did our participants. This risks glossing over the complexity of individual experiences, but we feel that this broader approach has value as a complement to more nuanced, in-depth studies of particular policies. We present this paper as novel empirical data about how different evaluators, researchers and policymakers understand and grapple with unintended consequences, not as a way of presenting a definitive account, but rather raising important methodological, theoretical, and normative questions about evaluation. With our participants, we have shaped some potential responses to these questions, which we hope will inspire thoughtful response in the wider evaluation community.

Results:

Participants acknowledged that unintended effects were common, and came about for a range of reasons including flawed policy design and implementation, unclear articulation of policy mechanisms or goals, or unclear or inappropriate evidence use, including evaluation techniques. Participants identified three main challenges: being able to identify and evaluate unintended effects, ensuring that the evaluation techniques themselves do not create the unintended effects, and being able to explain these effects. Participants also discussed possible solutions, including better use of theory, stakeholder engagement, and use of evidence. Perhaps unavoidably, the discussion tended to focus on unwanted, or otherwise negative unintended consequences. These may be more salient

to policymakers, or at least there is a more urgent desire to avoid them. However, both positive and negative effects can be unintended, and therefore the discussion below identifies both types.

Table 1: Summary of discussion on evaluation of UCs

	Practical issues	**Political issues**	**Ethical issues**
Policy evaluation - general issues	Evaluation part of policy – need to produce impacts leads to outcome-focused policies Evaluations not done, done badly, or ignored	Findings often not managed well or certain narratives prioritised	Not always clear who does evaluations, why, or for what purpose.
Challenge 1: Identifying UCs	No regulator or reporting system for UCs of social / public policy Research funding system also militates against capture of UCs	Lack of willingness to hear about UCs Political pressure to use particular methods, or select particular outcome	Not everyone equally able to claim harms or make unintended consequences of policies widely known
Challenge 2: Evaluating UCs	Evaluation methods not well suited to identifying UCs Selected methods may not be optimal choice to capture the full range of effects and processes	Asking for an evaluation gives the illusion of control 'Ownership' of and responsibility for policy evaluation	Evaluation audience may dictate scope and scale of evaluation, missing key indicators and outcomes
Challenge 3: Explaining UCs	Theory behind policy rarely articulated; UCs hard to explain.	Policy narratives not always explicit, or may have different public / private versions	Arguably unethical to commit resources where mechanisms of policy impact are not well understood

The evaluation of policy

Inevitably, the discussion also covered participants' experiences with policy evaluation in general. The intertwined nature of evaluation and policy was emphasised, as well as the political character of choices about evaluation. In particular, several participants suggested that a policy culture where measurable impacts are strongly emphasised will shape what policies and interventions are likely to be implemented.

Participants also suggested that the term 'evaluation' is highly polysemous. For example, one participant distinguished between a technocratic sense of evaluation ('what works') and a political sense (the 'success' of a policy as a function of the perceptions of other policymakers, the media or the public). Several participants emphasised the perceived need to maintain positive narratives about policies, which could lead to evaluation being ignored, added on as an afterthought, or poorly

designed (e.g. post-implementation). Partly in contrast, some participants drew a distinction between the maintenance of performative narratives about policy and the actual decisions and policy goals (e.g Nixon and Kissinger strengthening the narrative about aggressive bombing of Vietnam, while actually preparing to withdraw). Such public narratives may shape the statements made by elected politicians while having limited impact on their actual decisions. For example, ministers may change policy in the light of information about UCs, but not be able to publicly say so for fear of being accused of weakness or 'U-turns'. In general, a public line may run in parallel with a separate policy development process.

All these factors may shape how policies are understood in terms of their intended goals or outcomes, but considering UCs as a focus of evaluation further highlights the constraints and tensions involved.

Challenge 1: Being able to identify unintended effects

Workshop participants were split over whether UCs could be identified. Some participants felt it would never be practically possible to track every potential out-of-scope effect, whereas others felt that with improved policy testing and evaluation methods, UCs could be not only identified but addressed. However political constraints could mean that it was often considered unwise to seek out unexpected effects, as it could give the appearance of or raise uncertainty about the policy direction. More specifically, there seemed to be a link between the methodological and political challenges, in the sense that evaluating UCs often requires anticipating which UCs might be likely to occur. While this is also true of intended effects, the choices in that case are less likely to be controversial, since they translate the 'official' narrative of the intervention. In the case of UCs these choices are harder to justify, and may be rejected as arbitrary or unmotivated, which in turn means that evaluators are more politically exposed. Thus, thinking about UCs tends to cast doubt on the idea that the choice of which outcomes to collect data on can ever be purely technical and politically neutral. Of course, this also raises ethical questions about who gets to identify and claim unintended or harmful effects, with those most vulnerable being least likely to wield this political power.

Several participants suggested that the framework of policy implementation is not ideal, and that in many policy fields, interventions usually do not undergo effectiveness and safety testing before being widely implemented. By contrast with clinical healthcare, most policy areas have no regulatory framework to collect and analyse reports of UCs. The identification and evaluation of UCs is usually *ad hoc* and unsystematic, and conducted within the context of commissioned research projects which may not adequately mitigate political pressures. Hence, UCs may often be missed.

Challenge 2: Evaluating unintended effects

Most participants agreed that usually, in practice, UCs were not evaluated. This was partly due to the (very real) practical issues, such as a lack of flexibility in current evaluation designs and systems. Some participants pointed to the challenges of interpreting mixed evaluation findings: evaluators may see mixed findings as contradictions to be resolved by improved methodology or interpretation, rather than opportunities to explore the complexity of intervention effects within a system, and to more fully characterise intervention processes. Evaluation methods can give rise to UCs through measurement techniques (Bjørnholt and Larsen, 2014) . For example, one participant described a police campaign to tackle drug-driving, which involved making more roadside stops. The aim was to reduce drug-driving, but the policy actually raised the number of arrests of organised criminals – a positive UC in itself, but partly an adverse UC in political terms, since it gave the impression of a worsening crime-rate.

There was a wide-ranging discussion about who 'owned' evaluations; in the sense of paid for controlled, acted on, and responsible for. Some felt that policymakers were themselves the 'stewards' of evaluation, even when not directly involved. However, policymakers' own sense of ownership of policies varies widely over time and between different policy actors, which may affect their resistance to evidence of adverse UCs, and whether or not UCs are reported. Some funders are tolerant of 'failure' as long as there is some learning, although the churn in policy staff may mitigate against this. The relevance of UCs may also depend on the political goals of those commissioning or using evaluation evidence: for example, whether evaluation is intended to inform decisions about disinvestment, or to defend a particular policy position. Evaluators themselves, even when formally independent, may learn to anticipate such political and practical challenges and de-emphasise unwanted findings in their reporting of evaluation data, in order to maintain positive relationships or access future funding opportunities. Indeed, evaluators may need a higher standard of proof to have the dialogue about UCs, particularly adverse UCs, than they do with positive intended effects.

Participants also recognised that evaluation (from choosing what to evaluate, selecting design and targets, to dissemination) was a valuing process – but it wasn't always clear exactly whose values were being operationalised. The people targeted by interventions, or those involved in delivering them 'on the ground,' may have different and complementary perspectives to those brought by policymakers or evaluators. Policies are sometimes designed on the basis of what can be measured, rather than on what can be changed. Indeed, goal-setting is a technical exercise and conducted by administrators or political elites – not by those affected by policies. Measuring is a political activity, framing and limiting, which are therefore not neutral or rational (Bjørnholt and Larsen, 2014), and can be acted on politically. McLean et al describe an initiative to increase the 'number of swims per square foot of pool area', where this outcome could hypothetically be achieved by closing pools (McLean et al., 2007) – an outcome which is absurd in a practical sense, although coherent within the logic of measurement.

Challenge 3: Explaining these effects

Participants argued that even where evidence of UCs was gathered, the explanatory power of evaluations to account for them was limited. The political factors discussed above which hinder the identification and reporting of UCs will also tend to limit attention to their causal explanation. However, there are also methodological challenges. Methods which rely on statistical hypothesis-testing such as RCTs can only incorporate a limited number of pre-defined outcomes, and hence are likely to miss many UCs, while retrospective data-mining approaches (as one participant put it: "identify a huge bucket of indicators and run clever statistical analyses") may be too diffuse to offer real insight. The lack of theory in driving evaluation questions, designs and methods was discussed by participants, and it was noted that no significant attention had yet been paid to how to develop theories of harm, or operationalise these into evaluation processes.

Participants felt that it was ethical to only commit resources to interventions where there is sound reason to believe that it targets mechanisms which have a realistic chance of bringing about positive change. Otherwise we risk directing scarce resources toward interventions which are negligible in their effects, or even negligent (Hawe 2015b, quoted in Moore 2017).

Possible solutions

This meant that there was a strong ethical and moral imperative to address the issue of UCs, as well as a methodological argument for incorporating as wide a range of perspectives as possible into evaluations. Involving people who are affected, and other stakeholders, in evaluation planning

(known as 'empowerment evaluation') but may facilitate better evaluation and buy-in to the outcomes of the evaluation (which may be particularly difficult if there is evidence of adverse UCs). However, some participants reported that these more inclusive forms of evaluation are seen as lacking rigour. How then should evaluators ad policymakers respond to identified unintended consequences, and what are the solutions to the challenges above?

Responding to UCs:

Adverse unexpected effects could lead to strong emotional reactions amongst policymakers (denial, grief, anger) when policies they had personally invested in had not gone as planned. Personal, not just organisational or institutional responses, were important to consider when thinking about the politics of UCs. Policies which were seen to have failed (unexpectedly or not) were difficult to manage. There was significant pressure to simply ignore inconvenient evaluation findings, and even to create data in response to this pressure.

Participants agreed that policymakers and evaluators should recognise that a suite of interventions is usually required to achieve sustainable change in an outcome, while accepting that trade-offs would need to be made and that researchers and implementers had different expectations and may need to agree to disagree. They also suggested that adequate understanding of UCs was more likely where evaluators were able to:

- define the output of evaluations in advance
- achieve a conversation about the overall story of an intervention or a policy – leading to a revision of the underlying theory
- list consequences, setting out the theory of change with stakeholders, and consider risks
- continually and iteratively revisit the theory of change throughout implementation

On a practical level, participants emphasised the need for flexibility in designing evaluations, for example by piloting interventions, incorporating greater deliberation at the design stage, including break points or get-out clauses to reduce sunk costs during implementation, or implementing adaptive methodologies which can take account of shifting policy priorities during the evaluation period. However, some participants felt that research funders and commissioners were a barrier to the uptake of such methods, since they tended to emphasise protocol-driven methodologies and to expect researchers to produce an *a priori* evaluation plan and stick to it. (However, it was recognised that there was variation in this respect, with some funders more open to adaptive methods.). In general, participants argued that understanding UCs requires a pluralistic approach both to methods for policy evaluation and to processes of policy implementation.

Solutions

1. Methods: Firstly, participants discussed how to improve evaluation design. This would mean requiring commitment and honesty from evaluation funders. For example, it should ideally be clear whether a given evaluation is intended to improve the details of programme implementation details, or to inform decisions about investment or disinvestment. Participants emphasised the importance of agreement that evaluation is about trying to build a model to better understand what is happening. Setting the evaluation question was seen as a crucial opportunity to get these things right, by discussing: what decisions they were taking, who the evaluation is aimed at, assessing what level of 'success' is good enough, and how to measure it.

The discussion on methods centred around a need to recognised that rigour is about transparency of processes, and not a quality of certain research designs (particularly the RCT). The over-reliance on

the RCT was seen as driving certain types of questions focusing on a limited range of measurable outcomes, which hampers evaluators' ability to collect data on UCs. More broadly, the language and culture of RCTs shape assumptions about policy goals and drivers which may not facilitate a deeper understanding of how interventions function in context. However, the RCT was recognised as a useful method to test experimental questions, although not easily adapted to changing circumstances, or good at addressing broader processes and hence developing a better understanding of UCs.

Participants discussed a range of methods which could potentially improve identification and explanation of UCs. Qualitative Comparative Analysis (QCA) was suggested for analysis of policy effects and policy design. QCA allows for the possibility of multiple causation and may help policymakers think about 'types' of people and possible responses which could inform implementation. Cost implications of different assemblages could be estimated.

Participants strongly agreed that process evaluation and process data should be much more prominent in evaluations, otherwise people end up with evidence that UC have occurred but no understanding of why. However, it was recognised that this is difficult and involves potentially contentious choices on the part of evaluators. As already noted, several participants felt that evaluations which incorporate a range of perspectives will be better equipped to engage with UC. Decisions about granularity of results, targets, outcomes and so on needed to be much more open and transparent. In general, evaluators and commissioners needed to be more open about the critical and non-critical dimensions of evaluation, and the political and social dimensions which were the most important.

2. Better engage with, develop and apply theory: Participants recommended several models of evaluation to address these questions, including realist evaluation, QCA, and Process Tracing. What these methods all have in common is an attempt to articulate the mechanisms underpinning social change, or the theory which explains it. Participants spent significant time discussing how to improve the development of theory to underpin policies, and to underpin the evaluation of these policies. Suggestions included increased piloting of policies and innovations to identify different sets of interactions; to test key assumptions in the theory of change, and to keep re-ranking assumptions; and to concurrently run and revise the policy design and evaluation. One practical suggestion was to make greater use of risk registers, which are already completed for all policies to identify key risks throughout the process. Initially hypothetical, each risk is given a likelihood and weighted. Mitigation plans can be written, capturing contrary views and allowing feedback loops. Over time, as these are revised, the risk register becomes less a blueprint and more of a living document capturing what occurs on the ground.

3. Work with relevant stakeholders to produce theory-based evaluations. Evaluations which incorporate a range of viewpoints are likely to be more useful, and acted on. The coproduction of theories could help evaluators and policymakers to better grasp why polices have the effects they do, and to be able to collect relevant evidence to document these. Bonell et al describe a process to formulate evaluations of harmful effects (Bonell et al., 2015):

1. Scrutinising the assumptions underpinning the theory for the intervention's (positive) effects
2. Identifying inputs to interventions, processes and mechanisms by which these components are meant to lead to outcomes
3. Reflecting on unintended interactions between the agency of stakeholders and the social structures which constrain them

4. Drawing on existing mid-range sociological and psychological theories

The next step is to translate this into an operationalisable evaluation framework, and most importantly, build theory to enable more effective prevention in the future. Our workshops also suggest that working with multiple stakeholders, through a managed process of collaboration, would help evaluators and researchers to identify key theories, components and agents. As Weiss put it, "As we gain deeper awareness of the complexities and interrelationships that maintain problem behaviour, perhaps we can develop coherent, integrated mutually supportive sets of activities, incentives, regulations, and rewards that represent a concerted attack and begin to deserve the title of policy" (Weiss 1993: p105).

Discussion

Policies and interventions can have unintended consequences, but unexpected effects are not routinely sought by evaluators. This matters, because policies could be operating in ways which we don't understand, and they could be harming populations. In addition, unintended effects are likely to be underreported, and evaluation plans and resources are usually too fixed to be able to pick them up. There are significant political, practical and ethical challenges around the evaluation of UCs which highlight the need to better understand the aims of policies, the mechanisms by which they work, and to improve the evaluation of policies.

The role of evaluation in the policy process

Exploring unintended consequences shows that methodological and political concerns play out in the practice of evaluation, and that evaluation plays an important role in the policy process itself. Weiss argued that evaluation was political in three ways: evaluations are conducted on policies which are the product of political processes, evaluation reports become one part of the evidence jigsaw for politicians, and finally, evaluation is political itself – by the questions it asks, the roles it imposes on evidence and on scientists, and by the statements it makes about legitimate policies and policy reform (Weiss, 1993). The existence of unintended consequences not only points up the gaps in the empirical knowledge generated by evaluation – particularly our understanding of why policies have the effects they do – but raises troubling ethical questions about how far this knowledge should be taken into account in formulating and implementing policy.

From a methodological perspective, selecting relevant outcomes and methods; deriving evaluation plans through implicit or explicitly-understood theories, and interpreting and acting on findings are methodological and political choices. The challenges of identifying unknown unknowns was recognised by participants, who suggested potential solutions including use of a large toolkit of evaluation approaches, and better use of theory in the development of evaluating planning. Being able to evaluate unintended as well as anticipated consequences would imply a truly holistic approach to evaluation. This could be achieved by inclusion of stakeholders, potentially in the co-production of theories of change, and using these to develop evaluation plans. Developing theories of change with stakeholders, using collaborative methods, is not a new suggestion (Connell et al., 1995). The Theories of Change Evaluation developed by Aspen outlines steps to agree a programme theory which "is acceptable to stakeholder because of its existing evidence based or because it seems likely to be true in a normative sense"(Blamey and Mackenzie, 2007: 443). It attempts to create community engagement and ownership of the programme and evolution through their collaborative process, but are concerned mainly with what to do (implementation theory). Stakeholders agree what will constitute success, and what the causal pathways would be.

On the other hand, Realist Evaluation (Pawson and Tilley, 1997) also proposes discussion with implementers to map out mid-level theories about how different intervention participants will be affected by the intervention, and to test these theories with a range of methodological techniques. This process implies that the evaluator should look at as broad a range of possible mechanisms as possible (Pawson, 2008). However, where these theories are unclear or conflicting, realist evaluation theory implicitly puts the evaluator in the position of adjudicating between them, a position which seems incompatible with seeing the evaluator as one among many political actors. More generally, it tends to imply that the reality of the intervention process is exhausted by the sum of implementers' and evaluators' theories – an assumption which leaves little room for UCs which may be unanticipated by *any* of the actors involved.

Theories of Change is therefore good at providing broad strategic learning implementation theory, (Blamey and Mackenzie, 2007) but can be beset with problems where programmes are unclear, or poorly implemented (Bauld 2005). Realist evaluation is good at understanding processes and micro interactions, but may not adequately reflect the complex and conflictual process involved in large-scale intervention processes. Yet, bringing elements of these approaches together might be a good way to generate evidence-informed policy, by ensuring that the theories of change which are subject to testing and refinement in the evaluation process are grounded in shared ownership of the programme or policy and broad agreement about its goals. This process may be challenging, as stakeholders have to examine their assumptions about how programmes work, but this kind of developmental evaluation can help to refine logic models. (Patton, 1994, 2010).

Additionally, this kind of responsive process would enable evaluators to deal with emergent outcomes (versus pre-identified), non-linearity, recursive loops and disproportionate outcomes, and alternative causal strands (Rogers, 2008); in other words, with the hallmarks of complexity as opposed to complicatedness (Rutter et al., 2017). Where the challenges of complicatedness derive from the interaction of multiple stakeholders with divergent expectations – which, as noted, is implicitly the focus of the mapping of programme theories in much realist evaluation – the challenges of complexity derive from the possibility of emergent events which transcend these expectations. Importantly, these may include positive UCs as well as adverse ones. Particularly where interventions are genuinely shared with implementers and stakeholders, they may benefit from the agency of the latter – for example in creatively adapting the intervention to the context, or in finding ways round unanticipated obstacles – in ways which cannot in principle be anticipated by linear causal theories. As suggested by Hirschman (1967: 160-188; cf Lepenies, 2008), UCs may be a resource as much as a threat. Methods which allow evaluators or practitioners to identify and manage risks, which are at least partly foreseeable and quantifiable (such as risk registers), may not be able to adequately deal with the radical uncertainty to which truly complex systems are subject.

Ethical and normative issues

This also underlines that the ethical and political dimensions of evaluation cannot be separated from questions of methodology (Stame, 2018). The goal of adequately accounting for UCs in complex systems is bound up with the project of making evaluations more democratic: independent, not accepting of contingencies on their activities, and promoting democratic ideals (House, 2015). Weiss argued that few evaluations had "had a noticeable effect on the making and remaking of public policy" (Weiss, 1993: 98) , a view shared by some of the participants in the present study. This is a stringent test for evaluations, and probably mischaracterises policy change as a top-down process. Rather, if we share the hope that evaluations can affect public decision-making (Bjørnholt and Larsen, 2014; Dahler-Larsen, 2011), even if non-linearly, shared policy aims, and shared evaluations would help resources to be distributed more effectively. That is, giving up the claim to

epistemological mastery does not reduce evaluators' ethical responsibility, but if anything makes it more demanding (Schwandt 2018). Clearly, the solutions proposed by ourselves and our participants to the challenges of evaluation are not available to all evaluators, nor would all evaluators feel they are appropriate or necessary. Evaluators are motivated by diverse interests and hold different views about their role in the policy and practice arena. In this paper, we have described some of the normative and ethical challenges which are uncovered by examining unintended consequences. Much more attention must be given to developing a diverse set of responses to these challenges. Our proposals above regarding 'holistic evaluation' are a first step towards one possible response, but many others are possible.

For those with an interest in increasing the use of evidence in policymaking, this study has some clear implications. Firstly, we must recognise that researchers, practitioners and implementers learn through evaluation, as well as the policymakers. (Hawe, 2015a). Next, as we know, policymakers are most likely to act on evidence which is useful to them. But beyond 'another case of 'policy-based evidence', this illustrates the importance of shared criteria for credible evidence; shared assumptions, share belief in the process, and shared ideas about what good and useful evidence looks like. As Weiss knew "evaluation result are not likely to be persuasive to those for whom other values have a higher priority" [pp.98) – so achieving agreement around values is essential. Additionally, policymakers rarely commit themselves to directions or specifics (Ettelt et al., 2015)

How, then, can these shared criteria be developed and adhered to? Our study suggests that making values and decisions clear, acknowledging tradeoffs, and thinking through agency of actors in a collaborative discussion would be a good start. (Porter, 2015) Helping all stakeholders use appropriate theory, demonstrating a "clear understanding of how the problem under consideration is created and sustained in context" (Moore and Evans, 2017) is only possible through genuine collaborative discussion. However, this is at present an open question. Some important avenues for further exploration have been identified by this study, including to what extent have scholars grappled with the reality of unintended consequences as experienced by policymakers; whether holistic evaluation should, or always does imply collaborative working with stakeholders, and to what extent joint inquiry is a plausible mode of evaluation (see, e.g. Prainsack et al., 2010); the role of evaluation in the policy process more generally, and finally the ethics and normative frameworks which govern our responses to unintended consequences.

Conclusions

Exploring the nature and challenges of unintended effects can shed new light on the challenges of, and possible solutions to, the evaluation of complex policy problems. Practically, evaluating UCs is challenging, as we have inflexible funding, a dearth of reliable data, and fixed protocols which enforces the measurement of outcomes we can (or are allowed to), which itself affects questions we ask (Parkhurst and Abeysinghe, 2016). Politically, challenges include the need for success narratives, the drive to act quickly rather than strategically, and the policy process itself. Ethically, the evaluation of UCs shows us that evaluation practices means making decisions about what to evaluate and how, and balancing pros and cons of policies is an ethical choice which is often ignored or side stepped.

Evaluation itself can create the appearance of unintended effects, yet the political environment can dictate methods used. And evaluations can often miss important changes in context, during process, or outside of main timeframe. Evaluations can be conducted to measure scale and scope of impact, to assess value for money, inform future planning, and ensure accountability. Methodological rigour is important, but ideally mixed methods should be used to address theory-informed questions. (ICAP

2010). One way to do this is to privilege stakeholder experiences in the coproduction of interventions and evaluations, which requires the active management of group dynamics and politics (Maini et al., 2018).

In summary, evaluation researchers have proposed a number of ways to improve evaluations in order to capture what interventions do and how they work: By using theory, involving stakeholders, and being adaptive. Yet, using the lens of unintended consequences, it is clear that a combination of these approaches is required to evaluate public health interventions and policies, in a way which will inform us about how social change occurs in complex systems.

References

Adab P, Pallan MJ, Lancashire ER, et al. (2018) Effectiveness of a childhood obesity prevention programme delivered through schools, targeting 6 and 7 year olds: cluster randomised controlled trial (WAVES study). *BMJ* 360. British Medical Journal Publishing Group: k211. DOI: 10.1136/bmj.k211.

Allen-Scott LK, Hatfield JM and McIntyre L (2014) A scoping review of unintended harm associated with public health interventions: Towards a typology and an understanding of underlying factors. *International Journal of Public Health*. Springer Basel. DOI: 10.1007/s00038-013-0526-6.

Bamberger M, Tarsilla M and Hesse-Biber S (2016) Why so many "rigorous" evaluations fail to identify unintended consequences of development programs: How mixed methods can contribute. *Evaluation and program planning* 55: 155–62. DOI: 10.1016/j.evalprogplan.2016.01.001.

Bevan G and Hood C (2006) What's Measured is What Matters:Targets and Gaming in the EnglishPublic Health Care System. *Public Administration* 84(3): 517–538. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9299.2006.00600.x/full (accessed 1 February 2018).

Bjørnholt B and Larsen F (2014) The politics of performance measurement: 'Evaluation use as mediator for politics'. *Evaluation* 20(4): 400–411. DOI: 10.1177/1356389014551485.

Blamey A and Mackenzie M (2007) Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges? *Evaluation* 13(4). Sage PublicationsSage UK: London, England: 439–455. DOI: 10.1177/1356389007082129.

Bonell C, Fletcher A, Morton M, et al. (2012) Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine* 75(12): 2299–2306. DOI: http://dx.doi.org/10.1016/j.socscimed.2012.08.032.

Bonell C, Jamal F, Melendez-Torres GJ, et al. (2015) 'Dark logic': Theorising the harmful consequences of public health interventions. *Journal of Epidemiology and Community Health* 69(1): 95–98. DOI: 10.1136/jech-2014-204671.

Byrne D and Callaghan G (2013) *Complexity theory and the social sciences: The state of the art*. *Complexity Theory and the Social Sciences: The State of the Art*. Routledge. DOI: 10.4324/9780203519585.

Clark AM (2013) What are the components of complex interventions in healthcare? Theorizing

approaches to parts, powers and the whole intervention. *Social Science and Medicine*. DOI: 10.1016/j.socscimed.2012.03.035.

Connell JP, Kubisch AC, Schorr LB, et al. (1995) *New approaches to evaluating community initiatives: Concepts, methods, and contexts*. *The Aspen Institute*. DOI: 10.1177/1356389003094007.

Dahler-Larsen P (2011) *The Evaluation Society*. DOI: 10.11126/stanford/9780804776929.001.0001.

Egan M, Bambra C, Petticrew M, et al. (2009) Reviewing evidence on complex social interventions: Appraising implementation in systematic reviews of the health effects of organisational-level workplace interventions. *Journal of Epidemiology and Community Health* 63(1): 4–11. DOI: 10.1136/jech.2007.071233.

Ettelt S, Mays N and Allen P (2015) Policy experiments: Investigating effectiveness or confirming direction? *Evaluation* 21(3). SAGE PublicationsSage UK: London, England: 292–307. DOI: 10.1177/1356389015590737.

Hawe P (2015a) Lessons from Complex Interventions to Improve Health. *Annual Review of Public Health* 36(1): 307–323. DOI: 10.1146/annurev-publhealth-031912-114421.

Hawe P (2015b) Minimal, negligible and negligent interventions. *Social Science and Medicine*. DOI: 10.1016/j.socscimed.2015.05.025.

Hirschman AO (1967) *Development Projects Observed.* Brookings Institution Press. Available at: https://books.google.co.uk/books?hl=en&lr=&id=oz8PBAAAQBAJ&oi=fnd&pg=PP1&dq=Hirschman,+A.O.+(1967).+Development+projects+observed.+Washington,+Brookings+Institution&ots=zEO7P-l-Xy&sig=zWnGFU5M2trOIl33X9Ycaq995Ew#v=onepage&q&f=false (accessed 14 February 2019).

House ER (2015) The politics of evaluation and the evaluation of politics. *Evaluation* 21(4). SAGE PublicationsSage UK: London, England: 492–496. DOI: 10.1177/1356389015605200.

Lepenies PH (2008) Possibilism: An Approach to Problem-Solving Derived from the Life and Work of Albert O. Hirschman. *Development and Change* 39(3). John Wiley & Sons, Ltd (10.1111): 437–459. DOI: 10.1111/j.1467-7660.2008.00487.x.

Loke YK, Price D and Herxheimer A (2007) Systematic reviews of adverse effects: framework for a structured approach. *BMC Medical Research Methodology* 7(1). BioMed Central: 32. DOI: 10.1186/1471-2288-7-32.

Lorenc T and Oliver K (2013) Adverse effects of public health interventions: a conceptual framework. *Journal of Epidemiology and Community Health* 68(3): 288–290. DOI: 10.1136/jech-2013-203118.

Maini R, Mounier-Jack S and Borghi J (2018) How to and how not to develop a theory of change to evaluate a complex intervention: reflections on an experience in the Democratic Republic of Congo. *BMJ Global Health* 3(1). BMJ Specialist Journals: e000617. DOI: 10.1136/bmjgh-2017-000617.

McLean I, Haubrich D and Gutiérrez-Romero R (2007) The Perils and Pitfalls of Performance Measurement: The CPA Regime for Local Authorities in England. *Public Money and Management* 27(2): 111–118. DOI: 10.1111/j.1467-9302.2007.00566.x.

Merton RK (1936) The Unanticipated Consequences of Purposive Social Action. *American Sociological Review* 1(6). American Sociological Association: 894. DOI: 10.2307/2084615.

Mittelmark MB (2014) Unintended effects in settings-based health promotion. *Scandinavian Journal of Public Health* 42(15_suppl). SAGE PublicationsSage UK: London, England: 17–24. DOI:

10.1177/1403494814545108.

Moore GF and Evans RE (2017) What theory, for whom and in which context? Reflections on the application of theory in the development and evaluation of complex population health interventions. *SSM - Population Health* 3. Elsevier: 132–135. DOI: 10.1016/J.SSMPH.2016.12.005.

Parkhurst JO and Abeysinghe S (2016) What Constitutes "Good" Evidence for Public Health and Social Policy-making? From Hierarchies to Appropriateness. *Social Epistemology* 30(5–6): 665–679. DOI: 10.1080/02691728.2016.1172365.

Patton M (2010) *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. Available at: https://books.google.co.uk/books?hl=en&lr=&id=gd_RvUbSWnsC&oi=fnd&pg=PR1&dq=patton+1994+developmental+evalaution&ots=pSZFCPqRRe&sig=UZQxY4nrFjI8NFHhEfGn3JLjh_Q (accessed 13 February 2018).

Patton MQ (1994) Developmental Evaluation. *American Journal of Evaluation* 15(3): 311–319. DOI: 10.1177/109821409401500312.

Pawson R (2008) Pawson Invisible Mechanisms. *evaluation Journal of Australia*.

Pawson R (2018) The Realist Foundations of Evidence-Based Medicine: A Review Essay. *Evaluation* 24(1). SAGE PublicationsSage UK: London, England: 42–50. DOI: 10.1177/1356389017746718.

Pawson R and Tilley N (1997) Realistic Evaluation. *The British Journal of Sociology* 49(September): 235. DOI: 10.2307/591330.

Porter S (2015) Realist evaluation: an immanent critique. *Nursing Philosophy* 16(4): 239–251. DOI: 10.1111/nup.12100.

Prainsack B, Svendsen MN, Koch L, et al. (2010) How do we collaborate? Social science researchers' experience of multidisciplinarity in biomedical settings. *BioSocieties* 5(2): 278–286. DOI: 10.1057/biosoc.2010.7.

Rogers PJ (2008) Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation* 14(1). Sage PublicationsSage UK: London, England: 29–48. DOI: 10.1177/1356389007084674.

Rutter H, Savona N, Glonti K, et al. (2017) The need for a complex systems model of evidence for public health. *The Lancet*. DOI: 10.1016/S0140-6736(17)31267-9.

Sanderson I (2000) Evaluation in Complex Policy Systems. *Evaluation* 6(4). Sage PublicationsSage CA: Thousand Oaks, CA: 433–454. DOI: 10.1177/13563890022209415.

Sanderson I (2002) Evaluation, policy learning and evidence-based policy making. *Public Administration* 80(1): 1–22. DOI: 10.1111/1467-9299.00292.

Stame N (2018) Strengthening the ethical expertise of evaluators. *Evaluation* 24(4). SAGE PublicationsSage UK: London, England: 438–451. DOI: 10.1177/1356389018804942.

Warren RL (1973) Comprehensive Planning and Coordination: Some Functional Aspects. *Social Problems* 20(3). Oxford University Press: 355–364. DOI: 10.2307/799599.

Weiss CH (1993) Where Politics and Evaluation Meet. *Evaluation Practice* 14(1): 93–106. DOI: 10.1177/109821409301400119.

Weiss CH (1998) Have we learned anything new about the use of evaluation? *American Journal of*

Evaluation 19(1). Sage PublicationsSage CA: Thousand Oaks, CA: 21–33. DOI: 10.1177/109821409801900103.