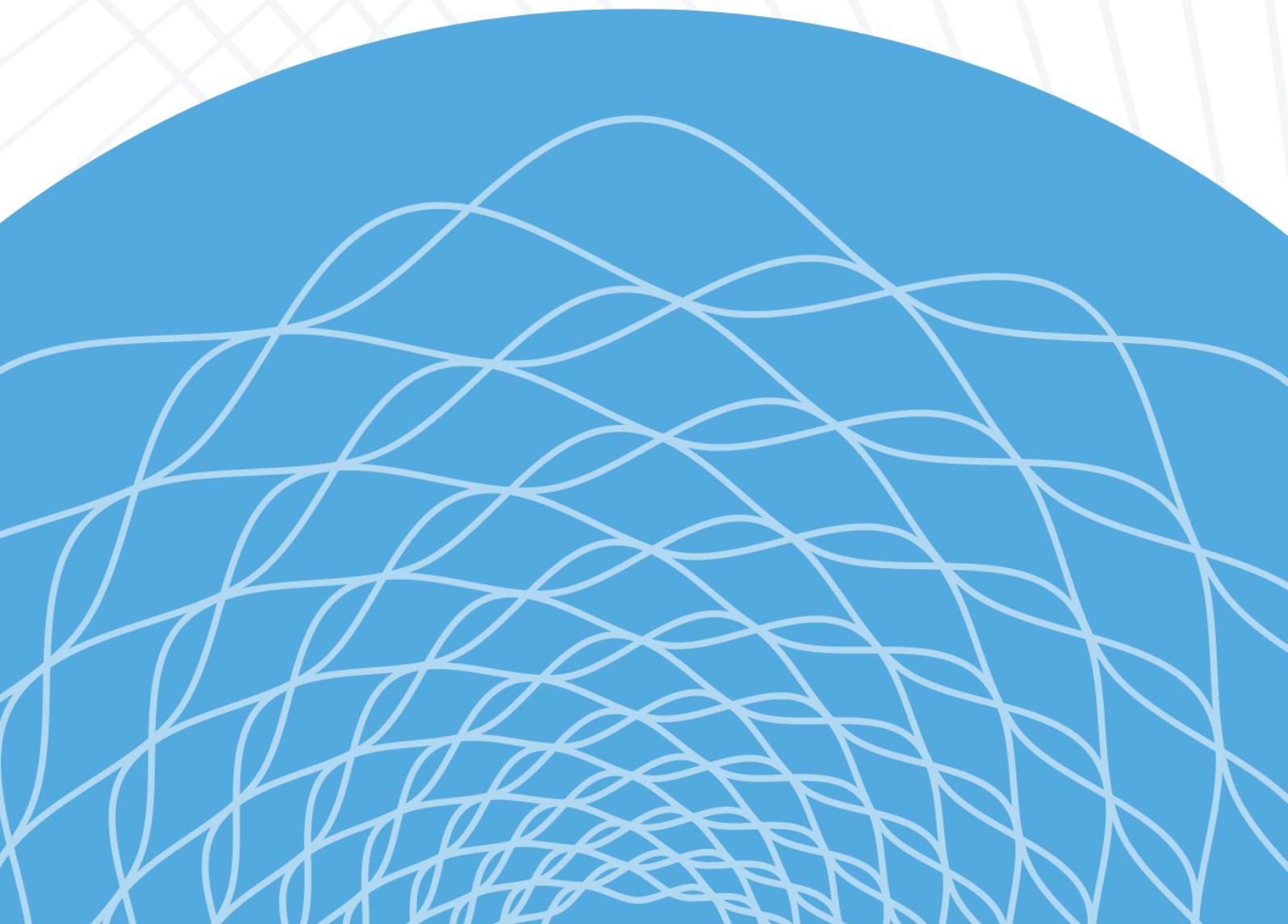


Bibliometrica®

Bridging design and evaluation: making
policy, programmes, and projects work

White Paper



HOW TO CITE

Bibliometrica White Paper [2022] *Bridging Design and Evaluation: Making Policy, Programmes, and Projects Work*

Copyright Bibliometrica® Ltd, United Kingdom, 2022

Downloadable from: <https://bibliometrica.org/>

ISBN: 978-1-7392458-0-1

DOI: 10.5281/zenodo.7753825

ABOUT BIBLIOMETRICA

Bibliometrica Ltd conducts analysis of policies, programmes, and projects that support scientific research and economic advance. It works around the whole policy cycle and draws on all the available evidence to support design and evaluation. It also advises on the change processes in the organisations responsible for policies, programmes, and projects. Its clients are governmental, research, and commercial organisations.

Contents

| | |
|---|-----------|
| HOW TO CITE | 5 |
| ABOUT BIBLIOMETRICA | 6 |
| 1. Introduction | 4 |
| 2. The Design Evaluation Divide | 4 |
| Design | 4 |
| Beginnings | 4 |
| Scepticism | 4 |
| A False Dawn for Understanding Causation | 5 |
| Design Remains Distanced from Causation | 5 |
| Evaluation's View | 6 |
| Avoiding Causation | 6 |
| Ignoring Weiss? | 6 |
| Independence of Evaluation as Field and Practice | 6 |
| Pragmatism | 6 |
| Outcome Focus | 7 |
| 3. Design-Evaluation Bridging | 7 |
| Some Awareness of the Other | 7 |
| Introduction | 7 |
| Evaluability Assessment | 7 |
| Programme Theory Based Approaches | 7 |
| Contribution Analysis | 8 |
| Regulatory Impact Assessment | 8 |
| 4. New Approaches | 8 |
| Methods | 8 |
| Structures and Conditions | 8 |
| An Effective Policy Cycle with Updating | 9 |
| The Bayesian Framework | 9 |
| The Challenge of Scale | 10 |
| Testing | 10 |
| Examples | 10 |
| Climate Change | 10 |
| Health and Social Care | 10 |
| Transport | 10 |
| Environment | 11 |
| Legal System / Legal Policy | 11 |
| International Development | 11 |
| The UK Position | 11 |
| Risk Reduction | 11 |
| Continual Learning | 11 |
| Policy Cycle and Causal Understanding | 12 |
| 5. Organisational Implications – Bridge Building | 12 |
| The Future | 12 |
| Causation as Bridge | 12 |
| Benefits of Bridge Building | 12 |
| Recommendations | 12 |
| Ensuring a Viable Policy Cycle | 12 |
| Requirement for Expertise | 12 |
| Building a Credible Evidence Base | 13 |
| Oversight | 13 |
| Gaming Behaviour | 13 |
| Scale Issues | 13 |
| Right to Challenge | 13 |
| 6. References | 14 |

1. Introduction

Policies, programmes, and projects ['interventions'¹¹] aim to change the way the world works. They need careful *design* to have a chance of effecting change, and thorough *evaluation* to find out not only what the effects were, but whether the design itself was responsible. And central to a design that works and to helpful evaluation is understanding *causation*.

The *design* of interventions and the *evaluation* should be connected by a common approach to causation. However, for much of the last sixty years, policy and programme design and policy and programme evaluation have been conducted as discrete tasks, attuned to different priorities, carried out by different people, using different techniques, and providing different results, ostensibly for the benefit of the same interests and rarely converging on the issue of causation. It might be said that policy and programme design and policy and programme evaluation are *Two Cultures*.

So those engaged in understanding how interventions are designed have often been unaware of consequences of design, while evaluation has sometimes ignored what design aimed to achieve. But there are now signs these two activities are connecting. New methods are emerging that focus upon causation and can make design and evaluation part of a holistic process. This bridge between design and evaluation is being built, but not before time.

There are a number of reasons why the construction of the bridge has been so slow. Conceptual difficulty is one. Another reason is that to investigate the mechanism / means by which policies and programmes take effect may challenge political ideologies. A further reason is that the separation of the activity of design from the activity of evaluation has been thought to make evaluation objective and independent, conditions important to developing it as a profession.

This white paper considers the reasons why the two activities have been pursued as separate enterprises. It then looks at why and how connections between them are now being made. The last section of the paper considers the next steps needed to bring these two activities together - what we term *design-evaluation bridging* - and what the implications are for organisations responsible for interventions.

2. The Design Evaluation Divide

"A house divided against itself, cannot stand."
[Abraham Lincoln]

Design Beginnings

The design of policies, programmes, and projects [design], is an activity central to the functioning of government. Undertaken usually by those responsible for the implementation of policy - conventionally a civil service/ state bureaucracy - policy design becomes a concern for government, a recognized area of academic study, and a priority for research by consultancies in recent times during the 1960s and 1970s.

In this inaugural phase, study of policy design gave significant attention to the aims of policies and programmes and how such aims emerged. But far less interest was paid to the means by which change was brought about. The tendency for policy design to avoid examination of how policies lead to change has continued, albeit with periodic realizations of a lack of focus upon this crucial issue.

One of the first to note the shortcomings of policy design was Salamon in the 1970s (1981, p. 256), noted in Howlett (2011, p. 4). He saw the focus on a comparative approach to programmes as a distraction from finding out what was really going on and why:

'The major shortcoming of current implementation research is that it focuses on the wrong unit of analysis, and the most important theoretical breakthrough would be to identify a more fruitful unit on which to focus analysis and research. In particular, rather than focusing on individual programs, as is now done, or even collections of programs grouped according to major "purpose," as is frequently proposed, the suggestion here is that we should concentrate instead on the generic tools of government action, on the "techniques" of social intervention.'

Scepticism

In parallel to attempts to construct an approach to understanding how and why policy worked was a strong sceptical strand of work, located within the field of implementation studies, a sub-field of policy studies. Highly influential in this area have been the contributions of Pressman and Wildavsky (1984) and also Lindblom (1959). They made a powerful sceptical challenge that questioned whether the designers of policy could ever realistically achieve their objectives. These and other authors argued that policies rarely functioned as intended, and that very often the

¹¹ The UK Treasury assumes a three-level hierarchy of governmental initiatives, policies being the highest, and then programmes, and then projects in descending order below it (HM Treasury, 2022). In common with a general usage, we refer to 'initiatives' and 'policies and programmes' synonymously.

means by which policies would take effect were not fully specified before implementation.

Pressman and Wildavsky's view of the difficulties of implementation centred on the limitations for agreement amongst different groups when complex decisions had to be made. Bowen considered this probabilistic approach overly pessimistic (Bowen, 1982) and sounded a note of optimism. However, Pressman and Wildavsky's perspective has remained a plausible picture of the challenges facing policy makers, and a dominant influence upon later writers. For example, Kingdon's sceptical critique of policy and programmes, (1984) suggested the rationales underpinning policies and programmes might be so diverse [a jumble of ideas, sometimes waiting for a policy entrepreneur] that they could not be translated into a functioning programme design.

Nelson's contribution to the debate on the complexity of social policy interventions drew a contrast between technical problems [sending a man to the moon] and dealing with complex social problems: the former was soluble with enough time and money, the latter, intractable, irrespective of how much time and effort were spent (Nelson, 1977).

A False Dawn for Understanding Causation

However, a focus on 'tools' and what could be achieved by programme design activities that focused on how change occurred gained momentum in the 1990s. This led to broader and more sophisticated attempts to understand how mixes of instruments would operate.

Increasing attention to 'instruments' by which effects were achieved, including consideration of suitable mixes of instruments, inspired later work outside the main policy design literature. In the innovation policy context the work of Flanagan, Uyarra, and Laranja (2011) reflects on the interaction between policy levels, between actors, the confusing effect of time, a position and which leads them to re-emphasize Lindblom's incrementalist position on what could achieve (Lindblom, 1959).

But a focus on tools was paralleled and, to some extent, gradually overshadowed by a more abstract and distanced perspective upon policy development. This trend led Howlett (2011) to claim there was a decline in the study of policy design *qua* basis for instrumental action. Howlett considered that policy studies had moved its focus away from policy as an instrument of government and towards the institutional frameworks in which policies arise. This change led away from understanding causation in specific cases.

The 'perspectival shift toward the latter' (Kooiman, 2000, p. 126) reflected a move from 'the managerialist state to a deliberative model' (Durant and Legge, 2006) quoted in Howlett (2011, p. 8). The

state and its capacity for agency was therefore viewed as circumscribed within relations defined by institutional frameworks. The state's scope of action was seen as restricted by globalisation, and with a consonant loss of authority at policy design / state level.

Howlett's claim has been challenged by Saetren (2015) who saw not decline but growth in the study of policy design. Later writers looking back with the benefit of hindsight see a gradual bifurcation in a comparatively vigorous and growing study of policy design, (Siddiki & Curley, 2022), with, on the one hand, a concern with how policy is formulated, and, on the other, a concern with its aims, including the means by which it achieves its effects.

Design Remains Distanced from Causation

Policy design, however, remains unwilling to tackle causation. Capano and Howlett (2021) reviewing their field of policy design recently note the continuing absence of work done on causal relations, i.e., the focus on why things work and when they do.

'Too often in the field of policy design, as in policy sciences and public policy more generally, explanations of 'what works when' are based on weak causation or a 'heuristic' framework, an often acknowledged to be unrealistic set of assumptions about irrational/rational behaviour, or a set of correlations between government actions and outcomes which are often mistaken for causes. Or they can be based on a causation 'derived,' from "what works" approaches, based on counterfactual estimates (Goertz and Mahoney, 2012; Heckman, 2005). Overall, an actual focus on realistic causation is often absent.' (Capano & Howlett, 2021, p. 142).

Their suggestion that policy design should employ a 'mechanistic approach' to help government operate effective tools (and initiatives) is the recent example policy studies' approach to the refractory problem of design.

'That is, a mechanistic approach to policy design strengthens decision-makers' analytical capacity by making it clearer what should be analysed and why. This approach then allows appropriate policy tools to be chosen to 'fit the job' and helps inform the calibration of those tools to ensure their effectiveness.' (Capano & Howlett, 2021, p. 142).

The authors continue, citing Weick:

'Paraphrasing Weick (1989), [we note] policy analysis should craft explicit hypotheses about the linkages between the input (design choices) and the output, including especially the specification of the process (the mechanistic causal chain) through which a policy design is

expected to lead to a specific outcome.’(Capano & Howlett, 2021, p. 143)

This mechanistic approach adopts a hierarchical and nested ontology inspired by critical realism of Bhaskar (Bhaskar, 2008), cited at (Capano & Howlett, 2021, p. at 146). A layered approach to the characterization of policy design simplifies a complex picture in which the causation that results in behaviour that policies have targeted is set within a hierarchy of mechanisms [first order ones² which nest within second order ones]. Second order mechanisms are considered to include both positive and negative effects, positive effects being a contribution to greater policy effectiveness and very often constituting policy learning.

The conceptual approach they define may distinguish the nested character in which policy design and operation occur, but this treatment of causation is abstract, seeing causation in the form of relations between different layers. The recent policy design literature therefore asserts causation to be a central theme and proposes the need for a mechanistically inspired approach to revealing it, but leaves explanation at a general and abstract level without elaboration of the specific methods for appreciating the magnitude of any effects.

Evaluation's View

Avoiding Causation

Evaluation as a field of professional activity and body of expert and self-critical knowledge has often downplayed causation in policies and programmes, and ‘worked away’ from the design of policies and programmes.

An early explanation of the distance put between policy and programme design on the one hand and evaluation on the other lies in the Post-war era of the Experimenting Society’s attempts to create change in very novel and uncertain contexts. The leading evaluation theorist of the time, Donald Campbell, acknowledged that society was making ‘innovative reforms’ and ‘exploratory innovation’ (Campbell, 2016, p. 223 and 224), and while in this era it was recognised that ‘Evaluation supports the experimenting society’ (Picciotto, 2020) [in reference to Campbell’s view], there was great uncertainty about the design of such programmes and how [and later, whether,] they would work.

Programme design was therefore conducted by way of experimentation, in a novel context, with novel methods and with the expectation of uncertain outcomes. It was not possible to specify the means with any greater certainty, and for some writers, even

the specification of ends was only possible after the fact (Dery, 1984).

Ignoring Weiss?

While policy and programme design and evaluation have developed distinct cultures, there were signs during the 1970s that evaluators should not ignore causal properties of programmes. But this very early work emphasizing the need to understand policies’ and programmes’ causal properties and in which Weiss was a pioneer (Weiss, 1972) was not noticed until much later (Worthen, 1996).

Independence of Evaluation as Field and Practice

A further contributing factor that de-emphasizes causation has been the need during the development of evaluation as an activity to establish its objective stance. Scriven’s comments (Scriven, 1997) on ‘transactional evaluation’ and ‘participatory design’ (Scriven, 1997, p. 483ff), and in support of ‘goal free’ evaluation, in which the evaluator – a team or an individual – ‘not only never talks to programme staff but does not even read the programme rationale documents’ (Scriven, 1997, p. 485) represents the clearest statement in the evaluation literature of evaluation’s attempt to achieve objectivity and impartiality, a position taken in part to help evaluation establish itself as a *professional activity* - at one time but now longer a goal of the community of evaluation practitioners. And as the policy areas in which evaluation engaged grew increasingly diverse, and the extent of policy experimentation so marked, it has become impossible for evaluation’s diverse practitioners to establish standard approaches that would have supported professionalization.

Pragmatism

Secondly, through the influence of another important practitioner, evaluation was conceived as a pragmatic activity, more of an art than a science, and ‘not cast into a single mold (sic)’ (L. Cronbach, J., 1982) cited in (Rossi & Freeman, 1993, p. 30) and emphasizing continual change in policies, programmes and contexts. However, it should be noted that Cronbach’s work on construct validity within psychological science did emphasize the link between different concepts within law-like relationships through ‘nomological nets’ (L. J. Cronbach & Meehl, 1955, pp. 290-294), his major influence has been one of pragmatic engagement with policies and programmes.

² (Capano & Howlett, 2021, p. 150) refer to first-order mechanisms as ‘those psychological and structural characteristics of policy actors which directly affect their behaviour and reaction to policy cues’. So called ‘second-order’ effects could on the authors general definition of the term include academic reflection and programme evaluation.

Outcome Focus

A further trend that sees evaluation 'work away' from design and causation is seen in the 1970s with the establishment in 1976 of the Evaluation Review (Parsons, 2001) where evaluation is increasingly conceived to have as its primary focus the examination of consequences – i.e., of outcomes. Evaluation is then confined to a single segment in Parsons' representation (Parsons, 2001, p. 546) of Palumbo's policy cycle and information cycle.

Thus, while evaluation contributes to design, it does so during formative evaluation, but only during policy and programme implementation, and not systematically and beforehand.

The principal emphasis in evaluation at this stage in its development moved towards a focus of evaluation as an activity of 'valuing' (Shadish, 1998). The study of causation and the question of what makes programmes work in practice becomes secondary. Causation appears in a list of key knowledges and pressing issues for evaluators only at point seven in Shadish's top ten (Shadish, 1998, p. 5).

Other important developments that have placed emphasis upon quantification and qualification of outcomes have arisen during the area of the New Public Management [the 'NPM'], an approach to the provision of public services developed in the UK and then exported world-wide (Hood, 1991). The NPM's approaches to the provisions of public services aimed to prioritize efficiency through competition. This required the development of organisational performance management systems based on indicators that could measure outcomes.

As Mawhood (1997) has noted, the approach had a strong emphasis upon outputs and not inputs³. NPM has had many outcomes, and has produced many surprises (Hood & Peters, 2004) but arguably its most significant result is the gradual enforcement of a outcomes culture in which measurement of impact or proxies for it via the use of indicators is central.

3. Design-Evaluation Bridging

'Only connect' [E.M. Forster]

Some Awareness of the Other

Introduction

While design and evaluation have been often pursued as separate cultures, recent developments have emerged demonstrating awareness of the link between what a policy or programme seeks to accomplish [aims], what will be done to achieve them [means] and what is then achieved [impacts].

Approaches to understanding causation developed within evaluation can be observed in the three related approaches of: i) evaluability assessment; ii) programme theories based approaches including theory of change and contribution analysis, and; iii) regulatory impact assessment. All of these approaches see some continuity between design and evaluation. These are discussed briefly below.

Evaluability Assessment

Evaluability Assessment is an approach to programme design that recognizes a tendency for policy and programme design to fail to anticipate the need for effective evaluation / impact assessment. It argues that at the design stage, policies and programmes should be made ready for evaluation. This is to be achieved by way of ensuring that policies and programmes have identifiable objectives, whether and to what extent stakeholders agree upon those objectives, whether methods are stated to achieve the objectives [a 'programme structure and resources', (Trevisan, 2007, p. 291)], and whether there are the means to assess the impacts.

Evaluability Assessment is most closely associated with Wholey (1979) cited in (Trevisan, 2007) and his group at the Urban Institute in Washington D.C.. They initially emphasized evaluability, but then expanded their focus to cover the use of feedback to effect programme improvement, a step reflecting Wholey's term for the approach describing an evolutionary approach which was *sequential purchase of information* or 'SPOI'.

Programme Theory Based Approaches

The introduction to the 20th Anniversary of the journal Evaluation in 2015 gives an overview of this growing focus in evaluation on the causal properties of programmes over the previous two decades (Stern, Saunders, & Stame, 2015). These approaches attributed to varying degrees, the capacity of *agency* to policies and programmes, and more importantly, they also aimed to recover, from evaluation [ex ante, including simulation, or ex post] the means by which policies and programmes take effect.

Within the broad approach are many different emphases, as Stern (2018) has noted. Broadly, theory based approaches, which are variously and usually described as 'theory-based, theory-driven and theory-informed evaluation' (Stern, 2018, p. 8) – the terms are often considered a synonymous - have important differences. Stern's paper considers the differences between them and we note the differences in the following way: a) approaches which seek to develop understanding of how policies and programmes work in a particular context would be termed theory-based;

³ One implication of which has been the occurrence of gaming behaviour associated with such systems elsewhere in the world [e.g., the Soviet Union] (18).

b) where there was specific emphasis upon developing an understanding of programmes through their contexts and mechanisms, the term 'realist evaluation' has been used; c) where the approach emphasises the testing of theories that have been identified as underpinning policies and programmes, the terms 'theories of change' and 'process tracing' are emphasized, as are the processes by which change happens.

Contribution Analysis

The term 'contribution analysis' is a more recent development amongst these approaches. Its main objective is to trace the impacts of policies and programmes back to causes in order to assess the extent to which interventions caused outcomes – 'the contribution'. Contribution analysis as distinct way of looking at policies and programmes was developed by Mayne (Mayne, 2001), in part to address what Mayne referred to as a 'changing culture of public administration' (Mayne, 2001, p. 1) where there was increasing pressure on administrations to assess the impact quantitatively, e.g., in 'taxpayers' dollars' (Mayne, 2001).

The approach drew attention to the need for examination of a 'results chain' (Mayne, 2001, p. 7 & ff.), a term with a close approximation in meaning to *logic chart* for a programme which would give programme managers and the public greater confidence that the *outcomes* [rather than outputs, which are more easily measured] actually derive from an initiative. The approach makes use of the concept of prior information and probabilistic causation, 'multiple lines of evidence' (Mayne, 2001, pp. 12, & ff.) in establishing causality although it does not explicitly draw on a Bayesian / statistical approaches to establishing credible causes. The notion of refining the results chain is however an updating approach to the evidence which underpins the causal properties of programmes. This can be seen to have similarities with the sequential purchase of information approach of evaluability assessment discussed above.

Contribution analysis is however an approach with limited usefulness in evaluation in that while it seeks to answer questions of attribution, it does not seek to assess the actual extent of outcomes quantitatively. Readers of the 2001 paper by Mayne seeing the use of the term 'extent' (2001, pp. 5, 9, 12, 14, 16, 19) might well conclude CA was being put forward as a means of quantifying impact, but subsequent development of the approach confirmed the absence of such an intention.

In some applications of the theory of change approach, the view is taken that it is the role of the evaluation to determine causation, and while the *ex ante* approach might offer a view of how a policy or programme would cause an impact, it would be the

responsibility of the evaluation itself to provide a final verdict (Government of Canada, 2021).

One can generalize on the development of theory based approaches to say that they represent an important change in the climate of evaluation, from a focus upon methods to a way which centres on the policy or programme itself. As Chen noted (H. T. Chen, 1994):

"The method or methods to be used is/are contingent on the nature of the evaluation model constructed for a particular program and evaluation situation.... Hence, the scope of theory-driven evaluations is much broader than that of method-driven evaluations and can serve as a comprehensive framework for dealing with various evaluation needs."

Regulatory Impact Assessment

A further approach that connects design with effect is Regulatory Impact Assessment ['RIA']. RIA is considered to be a systematic approach to understanding the effect of policy. Its influence is in part the result of its becoming a mandatory aspect of policy implementation (Claudio M. Radaelli & De Francesco, 2010), thereby contributing to a growing awareness of the importance of planning and understanding effects of intervention.

Nevertheless, while it focuses upon what the effects of policy might be on stakeholders, its emphasis is not upon developing an understanding of how policies achieve their effects, i.e., of causation, being an *ex-ante* form of evaluation. Indeed, as an *ex-ante* procedure it can be seen as leading to an avoidance of a focus on actual outcomes: 'hence the opportunity cost of *ex ante* analysis is given by the money that is not invested in *ex post* evaluation or in any type of assessment taking place after regulation decisions have been made' (Claudio M. Radaelli & De Francesco, 2010, p. 22). This leads to the conclusion that RIA for all its attention to impacts maintains the gap between design and evaluation.

Furthermore, without the development of epistemic communities to develop RIA as a result of an emphasis, particularly in Europe, on simple emulation of RIA approaches, for example the transferring in procedures rather [e.g., the standard cost model] as Claudio M. Radaelli (2009, p. 1152) have noted, the detailed modelling of the design of interventions and how they take effect has been silenced in RIA.

4. New Approaches

Methods

Structures and Conditions

While there has been increasing interest in theory-based approaches, the adoption of such methods in

policy design and evaluation practice has until recently drawn little from developments from more formal theorizing in philosophy of causation by Pearl (1995) (2009) [later popularized (Pearl & Mackenzie, 2018)]. The American Journal of Evaluation does not cite Pearl's 1995 paper until Keele's (2015) discussion of mediation, a central concept in the new modelling of causation that Pearl and others have developed.

The causal properties of policies and programmes have been taken account of at the design stage and at the evaluation stage over a number of decades, but the recognition of the importance of causal properties is now increasing in part because of the awareness of structural models of policy and programme causation that provide a better understanding of the strength of causal links and the dependencies between different activities.

These new methods emphasize that causal properties of interventions take place through structures of activities which can be traced down paths and webs of actions. Such causation can be modelled probabilistically using Bayesian Networks ['BNs'] most profitably, but also other techniques.

An Effective Policy Cycle with Updating

This structured approach to programme evaluation that attempts to provide a more rigorous approach to causation and with it, to provide opportunities to update the policy or programme model through 'probative assessment of empirical evidence through Bayesian formalization' (Fontaine, 2020, p. 296) has been most extensively explored by Befani and Befani and others (Befani, 2020; Befani & Stedman-Bryce, 2017).

However, while the Bayesian updating approach is intended to handle uncertainty – challenges remain establishing the strength of causal links empirically (Befani, Elsenbroich, & Badham, 2021). The BN approach is a viable way of bridging between design and evaluation, although use on evidence-based methods to support its application must be made with care as such evidence is often single-sourced, and fails to draw upon the widest possible range of material (Cartwright, 2021).

The Bayesian Framework

The Bayesian framework for policy, programme and project design and evaluation however provides a realistic starting point to address the bridging problem. It does this through the specification of a functional model of an intervention that is both a formal design and evaluative framework. This tighter coupling of *ex ante* and *ex post* evaluation supports improved interventions through clearer and better understood design parameters and their effects, facilitates intervention updating [formative evaluation], openness and transparency, and the

development of an evidence base upon which further policy can be based. Using BNs in project design connects directly to evaluation *ex post*.

Functional models of interventions in BNs are constituted of sets of conditional dependencies located in decision nodes whose probability tables [NPTs] are set within a directed acyclic graph. The dependencies and probabilities of the NPTs [model parameters] can be obtained from empirical data and in their absence, from expert advice using elicitation techniques (Daniel Kahneman, Paul Slovic, & Tversky., 1982).

Programme design is in any context and at any scale beset by uncertainty. Bayesian modelling is inherently accommodating of uncertainty in that the probabilities of nodes can be modelled not only as labelled, Boolean, or ranked, but as continuous variables. Furthermore, the methods allow for the use of techniques to identify whether and how further information might address the absence of information associated with causal effects, and the cost benefit of acquiring it, see Wilson (2015) on value of information analysis [VOI]. Where uncertainty has a particularly concerning effect upon outputs and outcomes, sensitivity analysis can be used to identify key causal properties in the model to reduce risk on a node by node basis. So-called Tornado graphs visually represent the effect of variation in the target node of any choice of and combination of input or source nodes. BNs can be used not only in a run-forward mode but in a backward reasoning mode to estimate the programme inputs required to achieve levels of outcome and impact.

BNs have heavy computational requirements but two recent improvements have been made to reduce the complexity of calculating probabilities: the junction tree algorithm has simplified the computation of the conditional dependencies over larger more complex networks; and object oriented modelling has been shown to provide a way of reducing model complexity.

The approach in principle gives greater confidence to those designing interventions and increases confidence of those affected by them. Bringing model/intervention assumptions to the surface supports the goal of open government, particularly when, presented in parallel, the programme's logic of key assumptions is presented in a narrative account. 'In silico' assessment of BNs carries the additional benefit of identifying potential savings in programme design from scaling back intervention size.

BNs contribute therefore to improved intervention performance and cost-effectiveness; and the greater visibility given to the assumptions upon which intervention is based supports the goal of opening policy making to outside scrutiny.

Intervention management routinely employs KPIs to measure output and outcome performance. A corollary of having a clear intervention model is that KPIs can also be used to rate or measure the evidential strength of the assumptions of the programme.

The Challenge of Scale

In addition to the challenge of understanding the causal properties of government initiatives set within complex webs of relations [the 'policy mix'] which can include other policies and also extraneous factors not part of formal policy processes, is the increasing scale at which interventions take place.

The UK Government (HM Treasury, 2022) considers the scale of government initiatives to run from *policies* at the top level, through *programmes* at a middle level to *projects* at the smallest scale, with a further assumption on the relationship between the different scales that initiatives at a higher level are an aggregation of initiatives at the lower level.

On the assumption that models of causation and their effects can be aggregated, the attempt to understand the complexity of large programmes should be addressed by working initially at the smaller scale and assembling an understanding 'from the ground upwards' to obtain an overall assessment of the causal power and likely impact of the initiative. While this approach is hampered by the possibility of interaction effects between the elements [i.e., *endogeneity*], this should be the start of any attempt at modelling of the cause and effect of initiatives, with the next step an iteration from observing the initiative in action through evaluation and then back to the models to update them.

Testing

In a small number of cases, the design of a policy or programme including its causal properties is now being considered during evaluation, and evaluations are contributing to further design and redesign of programmes, thereby helping to update a policy or programme model of cause and effect. The state of the art shows a number of actual examples of policy and programme design and policy and programme evaluation from a range of fields that have used structural models including BNs.

In the next section, we refer to a number of these studies. We also refer to a small number of studies where the design-evaluation bridge has been built out of causal modelling, uniting design, and evaluation in a connected activity - finally building the *design-evaluation bridge*.

Examples

Climate Change

In climate change, van Sluisveld et al. (2017) documents a range of *ex ante* modelling undertaken by the UK in the development of low carbon strategies in the attempt to overcome serious shortcomings in previous regulatory impact assessments Russel and Turnpenny (2009).

Health and Social Care

Structural models are widely used in health and social care policy and programming. At a top level, the National Institute of Health and Care Research [NIHR] has begun to use causal modelling of its policy related research portfolio (Research, 2022). But examples of specific initiatives at smaller scales using a wide range of quantitative methods to establish causation include the study of disease prevalence forecasting and management in responses to SARS-CoV-2. da Silva, Ribeiro, Mariani, and Coelho (2020).

Ex ante evaluation in diet policy has been undertaken by (Dogbe & Revoredo-Giha, 2021) using structural methods which establish relationships between specific public health goals in terms of disease prevention, the pricing of foodstuffs, and impacts upon greenhouse gas emissions [GHG]. Vieux et al. (2020) also consider *ex ante* the sustainability of dietary modifications.

The study by Osman et al. (2020) has used structural models in evaluation on the effect of the 'nudge approach to public policy' in the context of organ donation. Their approach, which uses Bayesian network software Agenarisk⁴, provides a way of updating policy.

Transport

Transport project appraisal is another area in which *ex ante* evaluation of policies has been undertaken using structural models. The study by Geurs, Boon, and Van Wee (2009) shows that in the UK, project appraisal does not take into account many of the potential forms in which impact arises. The UK has an extensive model for project appraisal. As the authors note, 'It can be concluded that the UK transport appraisal guidance (WebTAG) includes a much broader spectrum of social impacts than the Dutch appraisal guidance (OEI), but it does not cover the full range of potential social impacts as identified in the literature.'

In traffic management (P. Chen & Zhou, 2016), traffic planning and user satisfaction has been modelled causally when intervention policies have been designed, important lessons emerging on the differences between methods used which have

⁴ [<https://www.agenarisk.com/>]

included structural models, in this case Bayesian networks in combination with regression models. In the rail transport context, (Chakraborty, Mengersen, Fidge, Ma, & Lassen, 2016) have used structural / Bayesian modelling.

Of particular note in the context of transport policy is the study conducted for the UK by Hutchinson and Pearson (2004) which represents an early use of causal modelling and the connection of *ex ante* and *ex post*. This study helpfully draws attention to the link between design and evaluation.

Environment

Work by Ulengin, Kabak, Onsel, Ulengin, and Aktas (2010) in the context of transportation and environment relationships notes the challenges faced by policy makers in developing policy responses in complex environments, and also observes that the earlier research by Richardson (2005) did establish relationships between factors but did not move towards establishing the causal relationship between them. Later work in this context by Ulengin, Onsel, Topcu, Aktas, and Kabak (2007) has attempted to make progress in this area of policy making by introducing structural [neural network and Bayesian] methods.

In land use policy, a number of studies stand out as applying this approach. Celio and Gret-Regamey (2016) created a Bayesian network complete with nodes and a conditional probability table created from survey and other methods, and a similar approach was also used for the Netherlands by (Landuyt et al., 2016). The study by Celio, Koellner, and Gret-Regarney (2014) is another ideal case of design-evaluation bridging with a Bayesian Network based causal model of policy effect updated with questionnaire responses from local actors.

Legal System / Legal Policy

In the field of legal policy, the study by Buseti and Vecchi (2018) outlines a process tracing route to theory building in *ex post* evaluation of a programme to improve the functioning of the Italian judicial process, but more significant is the much earlier study in the UK by researchers (Fenn & Rickman, 2011) working for Civil Justice Council who carried out a study to examine *ex ante* the effect of a new approach to legal fee charging in the case of low value road traffic accident claims, what became the Fixed Recoverable Costs Scheme.

This study was then used as a basis for the Department Constitutional Affairs in England and Wales's policy. The researchers reported that after

two years they were invited to review the scheme which their *ex ante* research had been instrumental in helping develop. This was a case where a prospective study led indirectly to the development of policy and where the prospective study was a basis for *ex post* evaluation of the policy, an ideal case of design evaluation bridging.

International Development

In the area of international development, the UK's Department for International Development (Department for International Development [DFID], 2014), begun to assert a central role of evaluation in policy and programme design nearly a decade ago:

It is only by ensuring that evaluations are relevant, of high quality and effectively communicated that DFID can create a culture in which there is high demand from senior management for evaluation and where it is integral to policy and programme design and implementation resulting in programmes that are continuously improved (Department for International Development [DFID], 2014, p. 5).

DFID's evaluation culture, which puts evaluation at the centre of policy design has had an effect upon the new department of state of which DFID became part in the merger with the Foreign and Commonwealth Office [the FCO] to create the Foreign Commonwealth and Development Office [FCDO].

Now the FCDO in bringing forward its first ever evaluation strategy retains the commitment made by one of its precursor departments of putting evaluation directly in the service of design, see FCDO Policy Paper on Evaluation Strategy, Outcome 3 (2022, p. 9).

The UK Position

Risk Reduction

In the UK, the Government's stated approach to policy and programme development⁵ and evaluation has evolved over the last two decades. Effort has been made to reduce risk in the design of interventions and in developing greater understanding of impact. These changes are now part of Government guidance in the Magenta Book (HM Treasury, 2020a) and in its Annex to the Magenta Book (HM Treasury, 2020b), the Aqua Book (HM Treasury, 2015), and Orange Book (HM Treasury & Government Finance Function, 2021).

Continual Learning

These key works of guidance emphasize 'a continual learning and improvement through learning and experience' (HM Government, 2020, pp. point E, page

⁵ The UK policy making framework uses the term 'appraisal' to refer to policy design.

6) to provide an evidence base that supports future programmes and policies. The Magenta Book in particular shows that the bridge between design and evaluation is being created: evaluation is now considered *before* [i.e. during design] and *after* an initiative is implemented (HM Treasury, 2020a, p. page 5); there is an emphasis upon feedback loops 'agile evaluation design' (HM Treasury, 2020a, p. page 12); a focus on shaping the design of an intervention, including the use of a theory of change to explain how the intervention would work *ex ante* (HM Treasury, 2020a, pp. pages 12, 20), and *ex post* (HM Treasury, 2020a, pp. pages 24-26).

Policy Cycle and Causal Understanding

Also significant in terms of introducing a design-evaluation bridge has been the further development in the later versions of the Green Book of concepts of *policy-cycle* and *process of change*. By encouraging a cyclical view of policy and programme making, and emphasising mechanisms whereby causes lead to effects, the UK Government's advice on the design⁶ of its initiatives now emphasizes the need for causal models in design that make explicit the logical change process ['change mechanisms'] which can then support policy learning. The contrast with the Green Book of 2003 (HM Treasury, 2003) is very clear.

The current guidance therefore considers the causal properties of programmes and projects, in particular in relation to the so-called SMART⁷ objectives, 'An explanation of the logical change process i.e., the chain of cause and effect whereby meeting the business needs will bring about the SMART objectives' (HM Treasury, 2022, p. 20), but currently, within the Green Book, this is at the strategic level only. In terms of the other dimensions in which a policy or programme operates, the need for specification and understanding of the *logical change processes* is not yet emphasized. The *design-evaluation gap* is still therefore a feature of policymaking.

5. Organisational Implications – Bridge Building

The Future

Causation as Bridge

Awareness is growing that it is important to connect design of policies and programmes directly with evaluation of impacts, and that this can be achieved by using models of causation as a bridge. A number of governments, including that of the UK, are open to the idea of putting causal models at the centre of design and evaluation. This activity should increase but without it, design and evaluation are likely to

remain separate spheres of activity, with government intervention less effective.

Benefits of Bridge Building

Bridge building makes a policy cycle more dynamic, producing a richer and more informed understanding of how impacts of programmes arise. It contributes to the goal of open government, making explicit the policy and programme assumptions that underpin the choices that governments make. Policies and programmes whose models of operation are better understood can be adapted during operation to save money and resources. This is vital in times of financial stringency.

While there are examples of where the design-evaluation bridge has been made, the connection of the two activities is not standard practice.

Government guidance on policy, programme and project design and evaluation does not yet apply causal modelling to all the contexts in which government initiatives operate. There is still a strong emphasis on outcome focused evaluation, and while outcomes are vital, so is finding the evidence of how impacts occur and using that to ensure future policies and programmes are more effective.

If the design-evaluation bridge is to be more widely adopted in the UK, how should this be done? This white paper has a number of recommendations.

Recommendations

Ensuring a Viable Policy Cycle

Policy, Programme and Project models should be clearly stated at design stage. Policies and programmes should have their model of operation assessed in evaluation and updated with evidence. This may be a complex task, and some field work in terms of an *ex ante* evaluation / appraisal may be required that exceeds current levels of policy or programme preparation. Structural modelling of policies and programmes can use a wide variety of methods including those grounded in Bayesian statistics to represent degrees of uncertainty.

Requirement for Expertise

The design evaluation bridge will not be achievable without the ability to combine different expertise. Policy making bodies across government depend upon enhanced capacities in design and evaluation and particularly in terms of raising their modelling expertise.

⁶ 'Appraisal' is the UK Government's term for policy / programme / project pre-assessment.

⁷ SMART stands for Specific Measurable Achievable Realistic and Time-limited, (HM Treasury, 2022, p. 19)

Building a Credible Evidence Base

Very often, policies or programmes will have very high levels of uncertainty about outcomes of, or along, the causal links in their model. Under these circumstances, supporting studies to reduce levels of uncertainty are warranted as they will ultimately contribute to understanding of likely policy or programme impact. Such additional work will increase costs of design, but causal modelling that reveals the dependencies between the activities in a programme may help those designing interventions to avoid unnecessary activities. *Value of information* studies can be used to assist this process by estimating the value of supporting studies.

Oversight

Policy and programme assumptions should be stated clearly and be made public. As the approach is complex, it is considered desirable that intervention models should be reviewed by an appropriate body. Before 2015, this role might have been taken by the Audit Commission. Today, the recommendation is that National Audit Office is likely to be the most suitable body to take the role of reviewer. It may be relevant to establish a new body, the Office of Programme Responsibility [OPR] to carry out such a task.

Gaming Behaviour

When there is a risk that publication of the explicit model of policy or programme operation heightens the risk of opportunistic behaviour by policy or programme participants, it is suggested that the causal modelling approach is used to identify gameable parameters. Such parameters should not however be publicly disclosed but used to test for the presence of gaming behaviour.

Scale Issues

A number of approaches have been adopted to deal with the varying scales of government interventions. We suggest a programme based attempt to model causality, which focuses on middle range causation. We suggest that when programme managers cannot specify their causal models at a certain threshold, more information is gathered to improve the likelihood of success of the intervention.

Right to Challenge

The openness of an approach which lays bare for public scrutiny the basis upon which government policies and programmes are intended to take effect [the 'why' of programme impact] is only realized fully when there is a right to challenge model assumptions on causation and impact.

6. References

- Befani, B. (2020). Quality of quality: A diagnostic approach to qualitative evaluation. *Evaluation*, 26(3), 333-349. doi:10.1177/1356389019898223
- Befani, B., Elsenbroich, C., & Badham, J. (2021). Diagnostic evaluation with simulated probabilities. *Evaluation*, 27(1), 102-115. doi:10.1177/1356389020980476
- Befani, B., & Stedman-Bryce, G. (2017). Process Tracing and Bayesian Updating for impact evaluation. *Evaluation*, 23(1), 42-60. doi:10.1177/1356389016654584
- Bhaskar, R. A. (2008). A Realist Theory of Science.
- Bowen, E. R. (1982). The Pressman-Wildavsky Paradox: Four Addenda or Why Models Based on Probability Theory Can Predict Implementation Success and Suggest Useful Tactical Advice for Implementers. *Journal of Public Policy*, 2(1), 1-21. Retrieved from <http://www.jstor.org/stable/3998208>
- Busetti, S., & Vecchi, G. (2018). Process tracing change management: the reform of the Italian judiciary. *International Journal of Public Sector Management*, 31(5), 566-582. doi:10.1108/ijpsm-06-2017-0158
- Campbell, D. T. (2016). Methods for the Experimenting Society. *Evaluation Practice*, 12(3), 223-260. doi:10.1177/109821409101200304
- Capano, G., & Howlett, M. (2021). Causal logics and mechanisms in policy design: How and why adopting a mechanistic perspective can improve policy design. *Public Policy and Administration*, 36(2), 141-162. doi:10.1177/0952076719827068
- Cartwright, N. (2021). Rigour versus the need for evidential diversity. *Synthese*, 199(5), 13095-13119. doi:10.1007/s11229-021-03368-1
- Celio, E., & Gret-Regamey, A. (2016). Understanding farmers' influence on land-use change using a participatory Bayesian network approach in a pre-Alpine region in Switzerland. *Journal of Environmental Planning and Management*, 59(11), 2079-2101. doi:10.1080/09640568.2015.1120713
- Celio, E., Koellner, T., & Gret-Regarney, A. (2014). Modeling land use decisions with Bayesian networks: Spatially explicit analysis of driving forces on land use change. *Environmental Modelling & Software*, 52, 222-233. doi:10.1016/j.envsoft.2013.10.014
- Chakraborty, S., Mengersen, K. L., Fidge, C. J., Ma, L., & Lassen, D. (2016). A Bayesian Network-based customer satisfaction model: a tool for management decisions in railway transport. *Decision Analytics*, 3, 1-24.
- Chen, H. T. (1994). *Theory-Driven Evaluations*. Newberry Park: Sage.
- Chen, P., & Zhou, J. P. (2016). Effects of the built environment on automobile-involved pedestrian crash frequency and risk. *Journal of Transport & Health*, 3(4), 448-456. doi:10.1016/j.jth.2016.06.008
- Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol Bull*, 52(4), 281-302. doi:10.1037/h0040957
- da Silva, R. G., Ribeiro, M. H. D., Mariani, V. C., & Coelho, L. D. (2020). Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. *Chaos Solitons & Fractals*, 139. doi:10.1016/j.chaos.2020.110027
- Daniel Kahneman, Paul Slovic, & Tversky, A. (Eds.). (1982). *Judgment under Uncertainty*: Cambridge University Press.
- Department for International Development [DFID]. (2014). DFID Evaluation Strategy 2014-2019
- Dery, D. (1984). *Problem Definition In Policy Analysis*: University of Kansas Press.
- Dogbe, W., & Revoredo-Giha, C. (2021). Nutritional and Environmental Assessment of Increasing the Content of Fruit and Vegetables in the UK Diet. *Sustainability*, 13(3), 23. doi:10.3390/su13031076
- Fenn, P., & Rickman, N. (2011). Fixing Lawyers' Fees Ex Ante: A Case Study in Policy and Empirical Legal Studies. *Journal of Empirical Legal Studies*, 8(3), 533+. doi:10.1111/j.1740-1461.2011.01219.x
- Flanagan, K., Uyarra, E., & Laranja, M. (2011). Reconceptualising the 'policy mix' for innovation. *Research Policy*, 40(5), 702-713. doi:10.1016/j.respol.2011.02.005
- Fontaine, G. (2020). The contribution of policy design to realist evaluation. *Evaluation*, 26(3), 296-314. doi:10.1177/1356389020902496
- Foreign, Commonwealth,, and Development Office,,. (2022). Policy paper FCDO evaluation strategy: Evaluation Unit, Economics and Evaluation Directorate. Retrieved from <https://www.gov.uk/government/publications/fcdo-evaluation-strategy/fcdo-evaluation-strategy>

- Geurs, K., Boon, W., & Van Wee, B. (2009). Social Impacts of Transport: Literature Review and the State of the Practice of Transport Appraisal in the Netherlands and the United Kingdom. *Transport Reviews*, 29(1), 69-90. doi:10.1080/01441640802130490
- Government of Canada. (2021). Theory-Based Approaches to Evaluation: Concepts and Practices. Retrieved from <https://www.canada.ca/en/treasury-board-secretariat/services/audit-evaluation/evaluation-government-canada/theory-based-approaches-evaluation-concepts-practices.html>
- HM Treasury. (2003). The Green Book.
- HM Treasury. (2015). The Aqua Book: guidance on producing quality analysis for government. In.
- HM Treasury. (2020a). *Magenta Book*. Retrieved from <https://www.gov.uk/government/publications/the-magenta-book>
- HM Treasury. (2020b). *Magenta Book Annex A analytical methods for use within an evaluation*. Retrieved from <https://www.gov.uk/government/publications/the-magenta-book>
- HM Treasury. (2022). *The Green Book: Central Government Guidance On Appraisal And Evaluation*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1063330/Green_Book_2022.pdf
- HM Treasury, & Government Finance Function. (2021). *Orange Book: Management of risk - Principles and Concepts*.
- Hood, C. (1991). A PUBLIC MANAGEMENT FOR ALL SEASONS. *Public Administration*, 69(1), 3-19. doi:10.1111/j.1467-9299.1991.tb00779.x
- Hood, C., & Peters, G. (2004). The middle aging of new public management: Into the age of paradox? *Journal of Public Administration Research and Theory*, 14(3), 267-282. doi:10.1093/jopart/muh019
- Howlett, M. (2011). *Revisiting Policy Design: The Rise and Fall (and Rebirth?) of Policy Design Studies*. Paper presented at the Paper Prepared for the General Conference of the European Consortium for Political Research (ECPR)
- Hutchinson, E. J., & Pearson, P. J. G. (2004). An evaluation of the environmental and health effects of vehicle exhaust catalysts in the United Kingdom. *Environmental Health Perspectives*, 112(2), 132-141. doi:10.1289/ehp.6349
- Keele, L. (2015). Causal Mediation Analysis: Warning! Assumptions Ahead. *American Journal of Evaluation*, 36(4), 500-513. doi:10.1177/1098214015594689
- Kingdon, J. W. (1984). *Agendas, Alternatives, and Public Policies*, Boston Little, Brown.
- Kooiman, J. (2000). Societal Governance: Levels, Models, and Orders of Social-Political Interaction. In J. Pierre (Ed.), *Debating Governance* (pp. 138-166). Oxford: Oxford University Press.
- Landuyt, D., Broekx, S., Engelen, G., Uljee, I., Van der Meulen, M., & Goethals, P. L. M. (2016). The importance of uncertainties in scenario analyses - A study on future ecosystem service delivery in Flanders. *Science of the Total Environment*, 553, 504-518. doi:10.1016/j.scitotenv.2016.02.098
- Lindblom, C. E. (1959). The Science of "Muddling Through". *Public Administration Review*, 19(2), 79-88. doi:10.2307/973677
- Mawhood, C. (1997). Performance Measurement in the United Kingdom. In Eleanor Chelimsky & William R. Shadish (Eds.), *Evaluation for the 21st century* (pp. 134-144). Thousand Oaks: Sage.
- Mayne, J. (2001). Addressing Attribution Through Contribution Analysis: Using Performance Measures Sensibly. *The Canadian Journal of Program Evaluation*, 16(1), 1-24.
- Nelson, R. R. (1977). *The Moon and the Ghetto*. New York: W. W. Norton & Company.
- Osman, M., McLachlan, S., Fenton, N., Neil, M., Lofstedt, R., & Meder, B. (2020). Learning from Behavioural Changes That Fail. *Trends Cogn Sci*, 24(12), 969-980. doi:10.1016/j.tics.2020.09.009
- Parsons, W. (2001). *Public Policy - An Introduction to the Theory and Practice of Policy Analysis* (1st ed.). Cheltenham: Edward Elgar.
- Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82(4). doi:10.2307/2337329
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.
- Pearl, J., & Mackenzie, D. (2018). *Th Book of Why*. London: Penguin, Random House.
- Pressman, J. L., & Wildavsky, A. (1984). *Implementation: How great expectations in Washington are dashed in Oakland: Or, why it's amazing that federal programs work at all, this being a saga of the economic development administration as told by two sympathetic observers who seek to build morals on a foundation of ruined hopes* (3rd ed.). Berkeley, CA: : University of California Press.
- Radaelli, C. M. (2009). Measuring policy learning: regulatory impact assessment in Europe. *Journal of European Public Policy*, 16(8), 1145-1164. doi:10.1080/13501760903332647
- Radaelli, C. M., & De Francesco, F. (2010). Regulatory Impact Assessment. In R. a. C. Baldwin, Martin and Lodge, Martin, eds. (Ed.), *The Oxford Handbook of Regulation* (pp. .279-230). Oxford: Oxford University Press.

- Research, N. I. f. H. (2022). Global Health Research Portfolio - Theory of Change. Retrieved from <https://www.nihr.ac.uk/documents/global-health-research-portfolio-theory-of-change/26036>
- Richardson, B. C. (2005). Sustainable transport: Analysis frameworks. *Journal of Transport Geography*, 13, 29-39.
- Rossi, P., H., & Freeman, H., E. (1993). *Evaluation: A Systematic Approach 5th Edition*. Newbury Park: Sage Publications Inc.
- Russel, D., & Turnpenny, J. (2009). The politics of sustainable development in UK government: what role for integrated policy appraisal? *Environment and Planning C-Government and Policy*, 27(2), 340-354. doi:10.1068/c0810j
- Saetren, H. (2015). Tales From the Crypt: The Rise and Fall (and Rebirth?) of Policy DesignA Rejoinder. *Administration & Society*, 47(9), 1134-1141. doi:10.1177/0095399715594646
- Salamon, L. (1981). Rethinking Public Management: Third Party Government and the Changing Forms of Government Action. *Public Policy and Administration*, 29(3), 255-275.
- Scriven, M. (1997). Truth and Objectivity in Evaluation. In E. C. W. R. Shadish (Ed.), *Evaluation for the 21st Century: A Handbook* (pp. pages 477-500). Thousand Oaks, California: Sage.
- Shadish, W. (1998). Evaluation theory is who we are. *The American Journal of Evaluation*, 19(1), 1-19. doi:10.1016/s1098-2140(99)80177-5
- Siddiki, S., & Curley, C. (2022). Conceptualising policy design in the policy process. *Policy & Politics*, 50. doi:10.1332/030557321X16346727541396
- Stern, E. (2018). What is Theory-Based Evaluation. Retrieved from https://ec.europa.eu/regional_policy/sources/docgener/evaluation/helpdesk/052018/6.Theory-based-evaluation.pdf
- Stern, E., Saunders, M., & Stame, N. (2015). Standing back and looking forward: Editors' reflections on the 20th Anniversary of Evaluation. *Evaluation*, 21(4), 380-390. doi:10.1177/1356389015608757
- Trevisan, M. S. (2007). Evaluability assessment from 1986 to 2006. *American Journal of Evaluation*, 28(3), 290-303. doi:10.1177/1098214007304589
- Ulengin, F., Kabak, O., Onsel, S., Ulengin, B., & Aktas, E. (2010). A problem-structuring model for analyzing transportation-environment relationships. *European Journal of Operational Research*, 200(3), 844-859. doi:10.1016/j.ejor.2009.01.023
- Ulengin, F., Onsel, S., Topcu, Y. I., Aktas, E., & Kabak, O. (2007). An integrated transportation decision support system for transportation policy decisions: The case of Turkey. *Transportation Research Part a-Policy and Practice*, 41(1), 80-97. doi:10.1016/j.tra.2006.05.010
- van Sluisveld, M. A. E., Hof, A. F., van Vuuren, D. P., Boot, P., Criqui, P., Matthes, F. C., . . . Watson, J. (2017). Low-carbon strategies towards 2050: Comparing ex-ante policy evaluation studies and national planning processes in Europe. *Environmental Science & Policy*, 78, 89-96. doi:10.1016/j.envsci.2017.08.022
- Vieux, F., Privet, L., Soler, L. G., Irz, X., Ferrari, M., Sette, S., . . . Darmon, N. (2020). More sustainable European diets based on self-selection do not require exclusion of entire categories of food. *Journal of Cleaner Production*, 248, 10. doi:10.1016/j.jclepro.2019.119298
- Weiss, C. H. (1972). *Evaluation research: Methods for assessing program effectiveness*. Englewood Cliffs, NJ: Prentice-Hall.
- Wholey, J. S. (1979). *Evaluation: Promise and performance*. Retrieved from Washington, DC:
- Wilson, E. C. F. (2015). A Practical Guide to Value of Information Analysis. *Pharmacoeconomics*, 33(2), 105-121. doi:10.1007/s40273-014-0219-x
- Worthen, B. R. (1996). The origins of theory-based evaluation. *Evaluation Practice*, 17(2), 169-&. Retrieved from <Go to ISI>://WOS:A1996WB29100009

