Academic Article

EJA EVALUATION Journal of Australasia

Evaluation Journal of Australasia
2024, Vol. 0(0) 1–22
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1035719X241270560
journals.sagepub.com/home/evj
Sage

Managing the risk of program failure: Propositional Evaluation as a tool for risk management

Andrew J Hawkins
ARTD Consultants & Charles Darwin University, Australia

Scott Bayley
Scott Bayley Evaluation Services, Australia

Abstract

Public policy and programs fail for a range of reasons, some preventable and some much harder to treat. Program administrators require methods for managing risk. This article argues that evaluation can be a useful tool for identifying and managing risk. It offers Propositional Evaluation as a cost-effective, risk-based approach to the evaluation of any policy, program, initiative, or simply, plan of action for the public good. Propositional Evaluation treats programs as propositions for action and focuses on developing sound, that is, valid and well-grounded propositions. It begins in the design phase and continues through delivery to provide a means for identifying and managing risks to achieving intended outcomes. This article situates Propositional Evaluation approach in the evaluation theory literature. It reflects on common causes of program failure and identifies those evaluators may help program administrators identify and mitigate. It provides a description of five critical and five residual risks that derive from the propositional approach. Evaluating these risks provides a structure for evaluation and adaptive management. The intended audience is program administrators, and evaluators working to support them who require prospective evaluation and adaptive management or rapid and cost-effective evaluation for program design and re-design.

Corresponding author:
Andrew J Hawkins, ARTD Consultants, 4/ 352 Kent Street, Sydney, NSW 2000, Australia.
Email: andrew.hawkins@artd.com.au

Keywords
evaluation theory, evaluation logic, adaptive management, risk management, Program Design Logic, Propositional Evaluation, program theory

What we already know

- Program evaluation is commonly considered a form of applied social science.
- Theory-driven evaluation sought to fix problems associated with a purely experimental or ‘black box’ applied social science by identifying the reasons and assumptions by which a program is intended to work.
- Constructivists and others advanced the idea of a hermeneutic-dialectic process of reasoning to make evaluative judgements; this approach to reasoning about the adequacy of program design has received less attention.
- Propositional Evaluation treats programs as logical propositions for action. Propositions will fail when they are logically invalid. They may also fail for other reasons.

The original contribution the article makes to theory and/or practice

- The article seeks to reinforce the concept of evaluation as a process of reasoning consistent with the logic of scientific discovery rather than with the generation or accumulation of knowledge in a scientific sense (positivist, realist, or constructivist).
- It situates Propositional Evaluation in the evaluation literature and articulates an explicit role for evaluation in identifying risks that a program may fail and managing those risks.
- It treats program evaluation as an ongoing process of risk management and adaptive management in complex adaptive systems.
- It provides a list of 10 risks that determine whether outcomes are likely to be achieved and the magnitude of outcomes relative to costs. This list will help those designing or evaluating programs to maximise their chances of success.

Introduction

This article posits that mostly, a program[1] is neither intrinsically effective nor ineffective. A program is a possibility for its participants to interpret and a course of action for its proponents to shape. Even with well-intentioned planning, rigorous logic, or broad theoretical foundations, a program is inherently risky. This vulnerability underscores the utility of a propositional approach to evaluation, shifting the emphasis from acquiring knowledge to managing risks associated with interventions into a volatile, unpredictable, complex, and ambiguous (VUCA) world (Bennett & Lemoine, 2014; European Commission, Joint Research Centre, 2021; McChrystal et al., 2015). This approach walks a middle path between rational analytical decision making and learning and adaption in response to emergent conditions (Mintzberg et al., 1998). It has a design phase based on reason and an implementation phase based on facts, and these constantly interact in the pursuit of sound policy and programs.

Part 1 of the article summarises the philosophical foundations of Propositional Evaluation, acknowledging parallel trends in evaluation theory. Part 2 shifts from theory to practice and raises some common causes of program failure. Part 3 shows how a propositional approach can be used to identify and mitigate 10 common risks to project success at any point in the policy or program life cycle.

In the article, we propose that programs should be treated as logical propositions and present a way of gathering information that concentrates on minimising the risk of failure. By doing this, we seek to enhance the effectiveness and influence of interventions in complex adaptive systems, and to provide a practical, action-oriented perspective for the field of evaluation. We also hope this article demonstrates the value of a logical approach to ex-ante evaluation during the design phase of any program. We are aware that for some people the terms ‘logic’ and ‘reason’ imply a positivist or mechanical approach to understanding change in the world, or assume we think human behaviour is driven by logic, or that we can plan a program and then just implement it. We do not think this way. A plan to intervene in the world is not the same thing as a hypothesis or causal explanation. Plans may not be true or false, but they should make sense.

If we are to persuade evaluation theorists of this propositional approach, it is important to have a sound philosophical basis. We align with a constructivist epistemology, emphasising a hermeneutic-dialectic process of interpretation and reasoning that values multiple realities, ethical considerations, and stakeholder perspectives to achieve a nuanced understanding of program value (Guba & Lincoln, 2005). But we do not share a constructivist ontology. Instead, we treat the essence, ontology, or fundamental nature of a program – insofar as evaluation and evidence is concerned – as a proposition (or argument) for collective action in the world. We embrace an ontological realist critique of causality in program design, acknowledging that change is driven by ‘mechanisms in context’ rather than by programs or activities themselves (Pawson & Tilley, 1997). Yet we do not focus on theory development, testing, or uncovering these stable or ‘transfactual’ (Bhaskar, 2014) causal relationships. We are concerned with the value of action in complex systems (Jackson, 2019; Kurtz & Snowden, 2003; Patton, 2019; Renger, 2015), ‘learning within the unknowable’ (Flood, 1999), and interactive planning that improves the future in the ancient and modern cybernetic tradition (ANU School of Cybernetics, 2022). Rather than seeking to diminish the role of social science and theory development in program design, we propose re-thinking their role. We envisage evaluators and program administrators as facilitators of practical outcomes. We see them as focussing on managing risks of failure rather than testing theory or uncovering universal truths about what works: thinking as engineers rather than as physicists.

In summary, we invite the reader to entertain the idea of programs as propositions or arguments for action and to treat them accordingly. Within this approach, program administrators and evaluators share a goal: a sound argument – that is, a valid and well-grounded (or well-supported) proposition – based on what they already know, and what they discover along the way. This 'scout' mindset (Galef, 2021) enables rapid evaluation similar to participatory action-research in a dynamic world where new conditions are constantly emerging.

Part 1. Background to the propositional approach

The propositional approach to program evaluation is inspired by logical and deliberative traditions, from Aristotle to contemporary theorists of reasoning such as Stephen Toulmin (1958) and Jürgen Habermas (1984) and philosophers of science and causality such as Roy Bhaskar (2014) and John Mackie (1974). In the program evaluation literature, the approach has parallels with the work of Guba and Lincoln (2005), Scriven (2008), Wholey (2011), and Schwandt (2015). It emphasises the creation of 'sound' programs through logical validity and empirical grounding, moving away from theory testing towards identifying risks in program design and managing the uncertainties of program implementation. This mindset fosters a collaborative, reasoned evaluation process that embraces multiple realities and perspectives. It champions pragmatism and reasoned action within complex systems.

The origins of this approach go back to Aristotle and the study of deductive logic, deliberative rhetoric, and the enthymeme. Drawing from the works of Aristotle, Hugh Lawson-Tancred (Aristotle, 1992) finds that despite the negative connotations resulting from its misuse, deliberative rhetoric is useful for reasoning about the value of uncertain collective action or finding the 'inherent persuasiveness' of that action – a topic with obvious connections to modern public policy in complex adaptive systems. The tool that Aristotle provided to apply in situations where conclusions are probable rather than certain, and when not all assumptions can be identified, was the enthymeme. This is intended to provide an argument structure or tool for logical reasoning about the value of action in complex and uncertain situations (Aristotle, 1992).

The enthymeme functions as a proposition for action rendered into an argument structure using a deductive syllogistic form (Cohen & Nagel, 1934). A proposition for action (or a program) can be understood as a series of premises leading to a conclusion. Each premise consists of a subject linked by a copula to a predicate – for example, 'Staff (the subject) intend (the copula) to apply the new guideline (the predicate)'. The subject should ideally be specified as precisely as possible to ensure greater accuracy. The premise and proposition can be rendered categorical by specifying which category of staff the condition applies to—for example, whether it includes all staff, only experienced staff, or another specific subset. The predicate may be a condition that is assumed (a necessary assumption) or a condition that derives from an activity of the program (an output that is considered necessary). Once all premises (assumptions and outputs) are selected and arranged, deliberation ensues as to whether the collection of premises warrants the conclusion (intended outcomes). That is, whether outcomes could logically be expected to follow if all prior premises were true, i.e. whether the outputs are delivered, and the assumptions hold. In this case, we anticipate that not all staff will intend to apply the new guidelines. At this stage, we may be uncertain which specific staff members are necessary for achieving our objectives and can update the premise as more information becomes available. However, if the collection of premises (assumed to be true) does not sufficiently support the conclusion, we would consider the proposition or program invalid. The difference between an enthymeme and deductive logic is that the premises in a proposition for action are uncertain, and many assumptions are unstated. This means those doing the reasoning are filling in the blanks and therefore creating risks, including those based on faulty assumptions and fallacious reasoning.

We have been asked how we developed the principles of Propositional Evaluation. It is beyond the scope of this article to fully explain the origins of Propositional Evaluation and the nature of the modified enthymeme (see Hawkins, 2020). In brief, the method was 'immanent critique' (Bhaskar, 2014). This involves analysing internal contradictions, tensions, and inconsistences within a purportedly logical or scientific account of program evaluation. We, like others, noticed much confusion among evaluators about the terms 'program logic', 'program theory', and 'theory of change' (Donaldson & Lipsey, 2006). What evaluators called 'program logic' lacked logical coherence. What they called 'theory' did not align with theory as understood in a scientific context. When evaluators talked about 'lower case t' theory, informal theory or 'theory incarnate' (Pawson, 2013), we felt they were referring to the 'reasons' and 'assumptions' that underpin a program (Weiss, 1995). In effect, we took Weiss' criticism of experimental evaluation as lacking explanation one step further. Not only did we use reasons to explain the basis of a program in social science, but we used the structure of a reasoned argument to provide a comprehensive account of the program's essential nature. This structure can be evaluated in the design phase, we can evaluate the logical validity of premises and conclusions, and during the delivery phase, we can evaluate the extent to which the program's premises manifest in the context in which the program is delivered. This is not actually uncommon in practice. A recent paper (Grey, 2024) reviewed the role of theory in a sample of evaluation reports and found theory most often appears in the form of propositions, mostly as post hoc explanations and questioning of assumptions. We simply seek to formalise this critique of assumptions and unrealistic expectations in the ex-ante phase as well as during rapid cycles of review. This means that Propositional Evaluation does not focus on theory testing or accumulating knowledge. It does not focus on discovering general and abstract answers to the question of what works or even 'what works, for whom, under what circumstances, and how' (Wong et al., 2016).

This immanent critique is 'emancipatory' because it helps overcome the constraints of a scientific account of evaluation. Evaluation is often characterised as applied social science. Expenditure on evaluation is usually justified because it serves the public good. For many, there is a link between evaluation and science due to the prestige of science 'operating much like religion in Europe of an earlier age' (Chalmers, 2013), rather than because of a clear match between the logic of inquiry in science and the determination of the merit, worth, or significance of action in the world. By instead presenting a program as a proposition, we promote 'communicative competence' (Habermas, 1984). That is, presenting a program as such allows people to debate the program design openly, transparently and rationally, and not be alienated by language that implies only social scientists or theoreticians can engage in this debate – something that is essential for a deliberative democracy. Humans are by nature able to reason, so evaluator's role may just be to guide that process. Research has shown that groups of humans with subject matter expertise who are engaged in 'unmotivated reasoning' (that is, honest attempts at finding an answer rather than seeking to justify a predetermined conclusion) or have the scout mindset are effective at identifying the logical flaws in arguments (Galef, 2021).

Our final impetus to devise this propositional approach was the posthumous advice of Donald Campbell (1984) in his answer to his own question, 'Can we be scientific in the applied social sciences?' He provided a four-point plan to achieve this. First, fund many local programs, including funds for program staff to do evaluate as they see fit, to 'debug' them, and only later quantify the impacts of 'proud' (i.e. well established rather than new) programs. Second, for these evaluations, award multiple contracts rather than do one evaluation, and ensure competitive re-analysis of results. Third, focus on technical appendices written for academics and do not pretend to do a scientific evaluation on a program that is better described as an allocation of funds than a coherent program. Fourth, avoid measuring outcomes as a tool of administrative control – focus on improvement rather than accountability. Campbell said, 'if you are convinced of the impossibility [of treating evaluation as science], it is your moral duty to publicly denounce the pseudo-science in which we inadvertently find ourselves engaged' (Campbell, 1984, p. 333). We accept that advice. Spending on evaluation should primarily improve decisions about action in the world rather than seek to increase knowledge about the world. That is a focus for science and a largely unrealised ambition for evaluation. We offer instead an account of evaluation that emphasises logic and values and that we think is compatible with the approach of Michael Scriven, who said 'evaluation is a branch of logic' (Scriven, 2008, p. 66).

Evaluators like Guba and Lincoln (2005) and Schwandt (2015) have emphasised the significance of reasoning within evaluation practices. Guba and Lincoln envisage the evaluator as a facilitator, eliciting stakeholder consensus through the hermeneutic-dialectic process. Schwandt, drawing on Toulmin (2003), advocates for evaluative reasoning grounded in argumentation, urging for evaluators to adopt a flexible and adaptive process that embraces diverse perspectives. At the other end of the argument stream, Adversary Evaluation has a focus, albeit more debate than dialectic, on a process of explicit reasoning about the value of an evaluand (Picciotto, 2019). Propositional Evaluation shares the love of transparent reasoning that these approaches emphasise. We make the further claim that the nature (or ontology) of the evaluand is itself an argument rather than a theory, no matter how 'incarnate' (Pawson, 2013), and so provides a structure for program development, ex-ante evaluation and democratic deliberation.

The prospective evaluation synthesis (Datta, 1990) is another approach that is similar to the propositional approach. It has seen increasing applications in recent times (see, for example, Salm et al., 2022) and is a type of ex-ante evaluation that focuses on the likely impact of a proposed policy or legislative change. As such, it is about evaluating the design of an intervention. Examples of this approach are hard to find, and no systematic method of applying it is available, but Pawson implemented it well in his famous detailed reasoning about whether a ban on smoking in cars would work (Pawson, 2013). Propositional Evaluation provides a structure for this kind of evaluation. It sets out all the conditions or premises considered necessary for program success, which may be arrayed and interrogated one by one. They can also be assessed holistically to determine whether, collectively, they appear to be sufficient for the intended outcome.

Practical parallels are also found in Joseph Wholey's framework for the sequential purchase of information, which prioritises the acquisition of timely, decision-relevant information over comprehensive knowledge accumulation (Shadish et al., 1991; Wholey, 1979). His four-phase evaluation framework, developed in Wholey (1979, 1987, 2004, 2011) and Smith (1989), includes evaluability assessment, rapid feedback evaluation, performance monitoring and impact evaluation. This framework advances a stepwise approach to evaluation, investing in further assessment only when the likely usefulness of the added information outweighs the costs of acquiring it (Hare & Guetterman, 2014).

Program evaluation that involves data collection takes time, but decision makers, executives, managers, and other stakeholders often cannot – or will not – wait. Auditors, for example, perform rapid assessments that typically focus on whether program activities comply with legislation or regulations. Rapid feedback evaluation goes beyond such assessments by producing additional products; in particular, designs for further additional evaluation work if that is warranted (Wholey, 1979). Rapid feedback evaluations are an extension of evaluability assessments and begin only after there is agreement on how a program is to be evaluated, including on goals for assessing, controlling, or enhancing important side effects. Rapid feedback evaluations then use evaluation synthesis, small sample studies, program data, site visits, and discussions with knowledgeable observers to (1) estimate program effectiveness and indicate the range of uncertainty in the estimates, (2) produce tested designs for more definitive evaluations in the future, and (3) further clarify the intended uses of evaluation.

The sequential purchase of information leads to cost-effective evaluations that support decision making by focussing us on what we really need to know, not what we would like to know. It also clearly aligns with Scriven's (1981) principle of cost-effective evaluation, which is that evaluation should not cost more than the value of information it delivers. Measuring a program's outcomes is not logical if its foundational design is not in place or, to use Campbell's term, the program is not yet proud (Campbell, 1984). Propositional Evaluation provides a structure to guide the rapid purchase of information by setting out the proposition for action. We now transition from theory to practice, beginning with a broad consideration of program failure, and then focusing on Propositional Evaluation and how it can be utilised as a tool for risk management.

Part 2. Common causes of program failure

Policies and programs fail to deliver their intended results for a wide array of reasons: some are within the control of program administrators and some are not.

We start with a list of nine *symptoms* of program failure. We then consider seven *causes* of these symptoms. We identify that evaluation may treat one of these causes. Finally, in Part 3, we identify 10 *risks* to program success, because such risks are at least partially under public servants' control.

Symptoms of program failure

The following list of nine common symptoms of failure in government policies and programs is informed by the public administration literature. It is also grounded in the collective experience of the authors, who together have more than 60 years of experience in evaluation and have led or managed several hundred evaluation projects.

1. **The production of goods and services is inadequate.** The program simply does not produce enough outputs to adequately service the target group.
2. **The program's goods and services are of insufficient quality.** The quality of program activities and deliverables is not fit for purpose and does not meet the target group's needs.
3. **The production process is inefficient.** The program is consuming too many resources for what it is producing and the outcomes it is achieving.
4. **The program is ineffective.** It fails to fulfil its intended purpose. For example, a program targeting unemployed youth fails to reduce youth unemployment.
5. **Clients are dissatisfied and are complaining.** Because they do not have market choice and purchasing power, users of government services will express their discontent rather than taking their business elsewhere.
6. **Staff are dissatisfied and are leaving the program.** Program staff who are not satisfied tend to leave programs. Troubled programs may have annual turnover rates greater than 20%, which negatively impacts service quality and corporate knowledge.
7. **There are problems with coordination, and interagency conflicts.** It is common for struggling programs to engage in conflicts with other jurisdictions/ organisations, and for their services to operate in isolation from potential service delivery partners.
8. **The program does not adapt or innovate.** If a program fails to respond to changing client needs or external circumstances, it tends to lose its relevance.
9. **Program reporting is inadequate.** This is often a present in a program that is struggling to demonstrate its relevance and achievements.

For a further discussion of these symptoms of failure, see Bayley et al. (2012), Brinkerhoff (1991), Larson (1980), Light (2014), and Nutt (2002). In contrast, Luetjens et al. (2019) offer examples and analysis of successful policies and programs.

Underlying causes of program failure

As a direct consequence of failing, programs often experience a lack of external political support from key stakeholders; adverse publicity in the media; and criticisms from watchdog bodies such as the Ombudsman, the Auditor General, and parliamentary committees. These determine whether there is a perceived need for change to improve performance, and they help to identify the source of the performance pain. Public sector managers benefit from understanding the deeper causes of failure and as well as who cares enough about the performance problem to be willing and able to address it. We suggest seven underlying causes of program failure for a program administrator to consider.

1. **The program's mandate is unclear.** The program lacks authority or there is a confusion of roles across agencies (Brinkerhoff, 1991; Larson, 1980; Luetjens et al., 2019).
2. **The program's structure is inadequate.** The program's structure does not fit its environment or its strategy (Larson, 1980; Light, 2014).
3. **Performance leadership is poor.** The leadership style is incongruent with the program and there is inadequate governance/accountability (Brinkerhoff, 1991; Light, 2014).
4. **The culture does not foster success.** The program's culture and incentives do not support a focus on continuous improvement and achieving results (Bayley et al., 2012; Light, 2014).
5. **Systems are not fit for purpose.** Organisational policies, systems, and processes fail to support effective program management and service delivery (Bayley, 2012, 2022; Brinkerhoff, 1991; Larson, 1980).
6. **Resources are inappropriate.** The level of resources (financial, physical, people, and technology) and operational capacity is inappropriate for the program's design and systems (Bayley et al., 2012; Brinkerhoff, 1991; Larson, 1980; Luetjens et al., 2019).
7. **The program's rationale is inadequate.** The program's logic is unsound, its assumptions are untenable, or does not fit the program's external environment (Bayley, 2022; Luetjens et al., 2019; Nutt, 2002).

These lists of symptoms and causes of program failure are non-exhaustive. We seek merely to remind readers that program success is difficult to achieve and that there are many opportunities to fail. In Part 3 of this article, we will focus on the seventh cause of program failure: an inadequate program rationale. This is the cause that the program administrator may address most easily, at least partially, with an evaluator's help. This is not to say that good evaluation cannot influence the six other causes of failure –especially on mandate, leadership, culture, and the demand for quality evaluation (Bayley, 2022) – it certainly can, but improvements in these areas will depend on many factors outside the control of a single program administrator.

Part 3. Using Propositional Evaluation to manage the risk of program failure

Propositional Evaluation focuses on making programs that work here and now, when our knowledge is partial, the world is changing, and the people running programs are fallible. It focuses on questions such as 'What makes *this* a good idea?' and 'How can we make *this* work?' At its core, it treats a program as an argument structure or proposition in the form of 'outputs (major premise) + assumptions (minor premise) = outcomes (conclusion)'. As set out by Hawkins (2020), Propositional Evaluation is based on a modified enthymeme that incorporates a theory of causality focused on necessary and sufficient conditions. The most precise form of a proposition or program includes categorical premises.

Propositional Evaluation does not expect people to know everything at the outset or perfectly rational human designers. Building complicated and complex interventions is not easy, and we often learn as we go, 'building the aeroplane as we fly it'. It does not mean we expect to 'plan the work' and then 'work the plan', but it does mean that we should identify obvious failures of logic (often unfounded assumptions) in the planning stage and treat them as early as possible.

The goal of Propositional Evaluation is to develop sound programs. A program is sound when its foundational proposition and operational conditions are both logically valid and empirically substantiated. Thus, there are two main steps in Propositional Evaluation. First, we determine whether a proposition for action is logically valid (that is, that the program makes sense). This is about the key conditions (outputs, outcomes, assumptions, and constraints) that are *necessary* for the plan to be *sufficient* for some outcome. These conditions set the focus for the second step: collecting empirical data to determine whether the proposition or conditions (including assumptions) are in fact well-grounded (i.e. conditions are manifesting in reality). When a proposition is both valid and well-grounded, we can call it sound – another phrase might be 'evidence-based'. Any deficiencies in the proposition exacerbate the risk of failure.

The unique value of this approach is twofold. First, it prevents overly optimistic program designs from being funded in the design phase. Second, it avoids inordinate expenditure on measuring outcomes a program could not logically achieve; instead, it enables program administrators and evaluators to focus on managing the risk of failure in a rapid and cost-effective way.

Propositional evaluation and the reach, effectiveness, adoption, implementation, and maintenance framework

One reviewer of this article suggested it would be useful to compare Propositional Evaluation with the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework developed by Glasgow et al. (1999). We think this is valuable. RE-AIM was originally designed to ensure consistent reporting of research results, then evolved into a method of translating research into practice. It then evolved again into the Practical, Robust Implementation, and Sustainability Model (Puerto RicoISM), an extension of the RE-AIM framework that emphasises the importance of considering a broad range of factors that influence the practicality, robustness, implementation, and sustainability of interventions in real-world settings. PRISM deals with dynamic contextual factors affecting the three AIM outcomes – adoption, implementation, and maintenance (see Meador et al., 2023).

Crucially, RE-AIM begins with an intervention that is *known* to be effective and requires all important premises or questions to be identified in five elements considered necessary for a sustainable intervention, typically in public health. These elements are reach and effectiveness with intended beneficiaries, adoption, implementation, and maintenance issues for those delivering the new intervention. These five elements provide useful concepts and may guide evaluation when the proposition is about translating a clinically effective intervention into practice – who did it reach, was it effective, how was it adopted, what were the issues for implementation, and does there appear to be a sustainable benefit? Propositional Evaluation, on the other hand, does not require that we start with an intervention known to be effective. It does not predetermine what questions should be asked. Where RE-AIM applies to the archetypal public health intervention, Propositional Evaluation can be used with *any* proposition for action. It does not require prior knowledge of efficacy or limit itself to five elements. The focus on the overall logic brings any 'weak link' in the proposition (such as unfounded assumptions, unrealistic expectations/and insufficient activity) into focus. For example, RE-AIM may ask an evaluator to focus on the reach of a breakfast program – were the right children receiving it? Propositional Evaluation provides guidance on whether the program is designed in such a way that it could and likely will reach the right children. Propositional Evaluation is a more flexible and deeper approach to evaluation than RE-AIM and is applicable to a broader range of actions, especially where there is no clinical evidence base on which to rely.

Adaptive management of 10 common risks to program success

The Canadian Council of Ministers of the Environment define risk management as 'a systematic approach to setting the best course of action under uncertainty by identifying, assessing, understanding, acting on, and communicating risk issues' (Canadian Council of Ministers of the Environment, 2021, p. 3). Risk management should be a central concern for proponents of evidence-based programs to address complex problems in a VUCA world, where greater maturity in risk management is associated with positive organisational performance (Hartono et al., 2019).

Here, we describe 10 manageable risks in the order in which they are most often able to be observed using a Propositional Evaluation approach (see Figure 1). We have derived these risks logically and from our experience, rather than sourcing them from the literature. The first set of five risks are most important, as they determine whether outcomes will be achieved – we refer to these as critical risks. The next five determine the magnitude of outcomes relative to costs – we refer to these as residual risks (see Figure 1). In the following section, we provide some examples applying this approach.

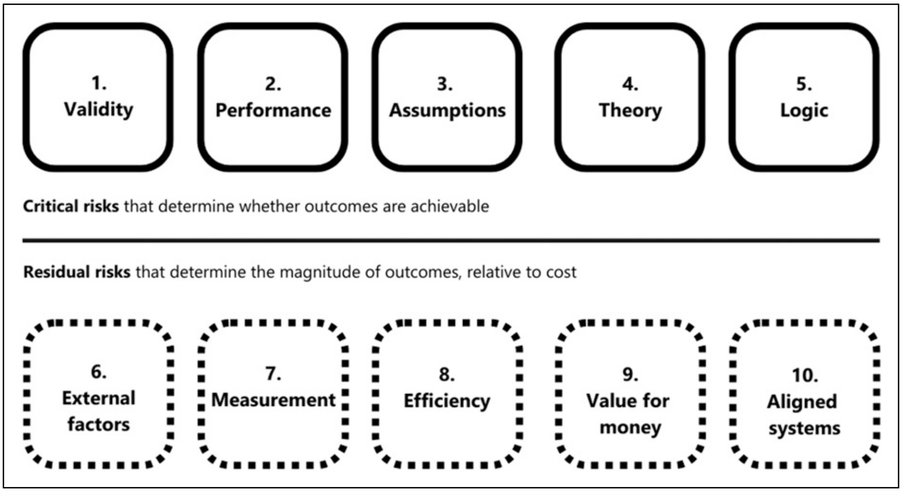
1.
Validity
2.
Performance
3.
Assumptions
4.
Theory
5.
Logic
Critical risks that determine whether outcomes are achievable
Residual risks that determine the magnitude of outcomes, relative to cost
6.
External factors
7.
Measurement
8.
Efficiency
9.
Value for money
10.
Aligned systems

Figure 1. Critical and residual risks to program success.

Validity (or design) risk. The intervention does not make sense on paper, heroic (highly unlikely) assumptions are being made, and our efforts could not reasonably be expected to achieve outcomes. Validity risks are quite common and occur when there are not enough reasons to accept the simple claim that outputs + assumptions = outcomes. Hawkins (2020) sets out how to determine whether a program is valid using the model of the enthymeme and the language of necessary and sufficient conditions in Program Design Logic. It is also possible to use the Toulmin (1958) model of argumentation to identify the strength of an argument (i.e. the reasons and warrants) that activities + assumptions = outputs. In either case, the logic of an argument or the validity of a proposition rather than a causal model is the standard by which we test the validity of design. Validity risk occurs most often where there is not enough honesty or humility about what a program could reasonably achieve.

Performance (or implementation) risk. It makes sense on paper, but we did not do what we said we would do, so it did not work. Performance or implementation risks are very common. They are also the simplest to understand and require relatively little explanation. They occur when certain activities are not implemented as intended – or without fidelity to a plan or guidelines about their delivery. Often this occurs when those tasked with delivery do not have the needed skills, time, incentive, or resources.

Assumption risk. It makes sense on paper, we did what we said we would, but assumptions about the operating context did not hold, so it did not work. Assumptions are conditions in the world that we are relying on, but program activities are doing nothing to bring about. Assumptions, just like outputs, operate as premises in a proposition. Reasonable assumptions support the validity of a proposition, heroic assumptions do not. We cannot list *all* assumptions that underpin action – they are limitless. The best we can do is identify those that we are relying on for the program to work, but about which we are reasonably uncertain. For example, if leaders implementing a new audit program must assume staff support the new way of auditing, but heavily suspect they do not, this would not be a reasonable assumption. If no new action is deemed possible, then this assumption is a constraint, and the program has created a validity risk. One of the simplest ways to enhance program design is to institute new activities so assumptions become the reasonable outputs of activities and are conditions that longer must be assumed. In the example above, it may require analysis of reasons staff don't support the new way and then new action that has a reasonable prospect of bringing about the condition that they do support the new way. Managing this risk in the design phase through discussion with subject matter experts is clearly faster and cheaper than waiting until the program is being delivered.

Theory risk. It makes sense on paper, we did what we said we would do, assumptions held, but necessary conditions (or outputs) did not materialise, so it did not work. Theory risk occurs when substantive social science theories (e.g. attachment theory or deterrence theory) as well as informal theories or reasons do not hold in the program context. Activities may work in some contexts but not others, or for some people and not others. This is a difficult risk to pinpoint, especially given the contested nature of the role of social science theory in evaluation (Donaldson & Lipsey, 2006; Grey, 2024). It may be that a genuine failure of theory is relatively rare, and what we encounter is an insufficiently nuanced theory or a failure to leverage theory. This happens when actions are not actually sufficient to generate outputs for all intended beneficiaries. Social science researchers and realist evaluators are well-placed to manage this risk by interrogating the contexts and circumstances in which purported casual mechanisms are fired by program activities. Managing this risk can be achieved by using condition statements that are more specific about the subject and the predicate, that is, for whom the output will be generated.

Logic risk. It makes sense on paper, we did what we said we would do, assumptions held, necessary conditions materialised, but the sufficient conditions (or direct outcomes) did not follow, so it did not work. This is the risk that the necessary conditions, once achieved, are not actually sufficient or adequate to achieve the outcome. Logic risk is related to but different to validity risks and theoretical risks. A logic risk occurs when despite everyone's best efforts at reasoning, the proposition is not logically valid. The premises in the argument (outputs and assumptions) may have manifested or been well-grounded, but this did not lead to the conclusion (outcomes). In technical terms, it is the risk that necessary conditions, once achieved, are not actually sufficient for the intended sufficient condition. That is, activities + assumptions ≠ outputs, or outputs + assumptions ≠ outcomes. This means that a logic risk is technically the same as a validity risk but is only apparent after the fact. If subject matter experts were perfect in their reasoning, then they would identify a logic risk as a validity risk in the design phase of a program. But no group of experts is perfect, and we all learn as we go. Logic risk is different to a theory risk because it is not about individual causal mechanisms we seek to leverage. It is about how outputs sum to an outcome.

The next five residual risks determine the magnitude of outcomes relative to costs. That is, they are risks that while the program might work in some way, it is not the best version of itself or the best option, or we do not really know its value.

External factor risk. It makes sense on paper, we did what we said we would do, assumptions held, necessary conditions (outputs) and sufficient conditions (direct outcomes) materialised, but longer-term contributory conditions (indirect outcomes) were not significantly impacted. External factors are by their nature outside the control of a program. In an open system, these will always exist. Strictly speaking, Propositional Evaluation and Program Design Logic are not concerned with mitigating external factor risk – it is concerned with ensuring that the program can achieve the outcomes or sufficient conditions for which it was funded. However, identifying external factors is important because its helps program funders see the gap between direct program outcomes (what the program will achieve) and indirect outcomes or contributions (why the program is funded) and be satisfied with the program's ambition. If funders do not consider the direct outcomes valuable enough, a Program Design Logic can pinpoint where additional action and resources are required. For example, if a homelessness outreach program is intended to be sufficient for ensuring people access the social housing system and establish a social housing tenancy, then it is the extent to which these conditions were achieved that should be evaluated. If it is intended to be sufficient for meeting housing need, it must deal with other major causes of homelessness, including poor mental health and unemployment. Propositional Evaluation does not support the utility of measuring a program's contributions to higher-level indirect outcomes or impacts *beyond those the program could directly achieve* – that is, outcomes for which activities were logically insufficient.

Measurement risk. The program is generating the intended outcomes, but the evaluation or monitoring system is not sensitive to the value being created, so the data provides an invalid or incorrect indicator of progress. Measurement risk is about the reliability and validity of the data that is being used to draw inferences about the program. This risk is unlike all the others in that it is not about the program per se, but about our efforts to understand its value. This is a risk that takes substantial evaluation expertise to manage. Reliance on evidence focused only on intended outcomes that can be quantified and monetised when the program is difficult to measure is a common cause of measurement risk. When a program is new and still working itself out, evaluators risk 'throwing the baby out with the bathwater' by measuring outcomes using experimental design. Drawing inferences from key performance indicators that bare little relation to a program's value is also a surprisingly common source of measurement error. Measurement risk can cut both ways – ineffective programs may be continued, and effective programs may be cancelled. Propositional Evaluation helps manage this risk by being humble about the specific conditions a program may be sufficient for achieving. To reiterate, the goal is testing the soundness of propositions, not testing hypotheses about stable cause and effect relationships, or quantifying attributable outcomes to program activity (internal validity) that may or may not repeat when similar activities are delivered in another time or place (external validity). To attempt this kind of measurement commits the fallacy of equivocating methods to determine the value or validity of an intervention into the world with methods to test the validity of scientific theories about stable cause and effect relationships in the world. Rather than post hoc measures of program impact, Propositional Evaluation prefers ex-ante (and ex-itinere) judgements of value relative to known risks, right now.

Efficiency risk. The program is generating the intended outcomes, but it appears that not every output (or necessary condition) is actually necessary – we may observe outcomes occuring even in the absense of certain actions, or outcomes without outputs we thought were necessary. This is the risk that certain conditions are superfluous or not required for the program to achieve (or be sufficient for) some outcome, and that the program is therefore inefficient. This may be difficult to identity ex-ante but can be determined post hoc using data where it becomes obvious that results followed in situations where certain actions or conditions were not enacted or seen to be effective. Qualitative Comparative Analysis is a method designed to identify necessary and sufficient conditions for an outcome (see, for example, Hanckel et al., 2021) and can be used in this situation.

Value for money risk. The program is generating the intended outcomes, but it is overly expensive, creating an opportunity cost when some other initiative may be more valuable. Value for money risk is easy to understand. It occurs when not enough attention is paid to how a public policy goal can be achieved using other mechanisms in the policy maker's tool kit, resulting in the costs of running the program outweighing the benefits it generates and low return on investment. Value for money risk can also occur if there is an efficiency or redundancy risk (or indeed if any of the other risks materialise). Methods to assess this risk are associated with economic and value for money analysis.

Aligned systems risk. The program is generating the intended outcomes but did not sufficiently identify how this would impact other parts of the system or whether the problems it is designed to address are just symptoms of a deeper 'root cause'. Aligned systems risk is last on this list for a reason. On one hand, no program should be developed without an understanding of the system into which it intervenes. On the other, it is almost impossible to avoid this risk in some form. A systems approach requires relationships with other actors who are, almost by definition, outside a program's control. There are two types of system where this risk can materialise: systems designed to solve a problem (e.g. housing and homelessness service systems) and systems that give rise to a problem (e.g. the socio-economic system that leads to homelessness).

To manage risk within the first type of system, a pragmatic approach is to focus initially on what the program is doing and to ensure it is done well. However, there is always a risk that the system needs something different to what the program is doing. A program may be effective and manage all nine risks listed above, but it may still fail to change system-level outcomes, simply because it is doing the wrong things right; that is, it is succeeding at doing something the system does not need. A program's activities may also be counter-productive and have negative impacts on systems. For example, a program designed to address trauma may place an unsustainable demand on trauma counsellors, leading to their exit from the system.

Within the second type of system, there is a risk that program outcomes address symptoms of a deeper root cause driving the problem. So, for example, an aligned systems risk may occur when a housing program is helping people establish tenancies, but untreated trauma is leading to unsustainable tenancies and repeat instances of homelessness.

Aligned systems risk tends to occur when the problem definition process is incomplete or when the scale of the program is not sufficiently wide (e.g. it does not include engagement with other partners) or deep (e.g. it does not consider root causes) for its aspirations to be realistic.

Using evaluation for adaptive management or risk management

In our experience, most programs fail due to straightforward instances of the first three risks described above – that is, outcomes were too ambitious and not logically possible based on the activities undertaken, the activities were not implemented with fidelity to the program plan, or the necessary assumptions on which the program relied were heroic and unlikely to be in place.

For example, consider a national regulator seeking to implement best practice regulation that promotes a culture of self-assurance or of maximising voluntary compliance. These are the right theoretical approaches, so such regulation may be unlikely to create theory risk. However, what if regulatory staff have low morale and are not sufficiently trained or incentivised to change their regulatory posture? This may be an assumption risk (usually it is assumed staff will at least try to do what they are asked) or a performance risk (training and the provision of resources are meant to ensure staff do what they are asked). Both of these risks can be identified and should be rectified *before* embarking on an expensive process to implement this change or measure outcomes.

Or consider the following example of using evaluation to support program risk management. The authors are aware of a developing country in which treasury officials are being trained in Australian methods for budget formulation and implementation. The logic of this program is expressed as follows: 'When budget officials have improved knowledge and skills, then the country will produce technically better and more sustainable budgets. This in turn will result in improved public services, increased foreign investment, greater economic growth, and improved public welfare in the long term'.

This logic contains three major risks. First, senior ministers and public sector officials are assumed to be committed to changing or improving the country's budgeting practices. This is assumption risk, as the program may have been promoted simply to secure increased international financial support. Second, it is assumed that the behaviour of budget officials is driven by their current level of knowledge and skills. This is theory risk. What if officials' skills and knowledge are not the key behavioural influences? For example, the country's budgeting policies and systems could be the source of the underperformance. Or perhaps the underperformance arises from self-serving elites controlling budget processes? Third, the reasoning that having a more sustainable national budget leads to better public services is a major leap. It is a validity or logic risk in that the program may not lead to the direct intended outcome and an aligned systems risk for an indirect or longer-term intended outcome or contribution. A change in budgeting practices may or may not result in greater funding being allocated to public sector agencies. In any case, additional funding may not be sufficient by itself to improve services to the public.

Undertaking a series of evaluations over time will help the program administrator to identify and manage the risk in this program. For example, an assessment of the program's design is likely to raise questions about the program's validity, as similar types of programs have been tried and failed in other countries. An evaluation of the program's implementation and intermediate outcomes is likely to find that the workplace behaviours of budget staff have not changed significantly and nor has the quality and sustainability of the country's budgets. Given the failure of the program's underlying strategy, there would be no need to undertake an impact evaluation. Each of these evaluation findings provides specific feedback to the program manager to better address the factors contributing to program failure.

Also consider the example of an Australian education program for farmers to reduce their use of water and fertiliser while maintaining on-farm outputs. By reducing their costs while maintaining levels of outputs, it was expected that profits would increase, and the environment would also benefit. An evaluation of the first two years of the pilot program in three locations showed disappointing results. Efforts to train the local farming communities were failing, and only a limited number of farmers were participating in the educational activities.

Program staff undertook a series of qualitative evaluative assessments to better understand what was happening. It turned out that each of the three areas had one to two lead farmers who influenced the farming practices in that community. Rather than training all farmers in an area, program staff needed to identify the local lead farmers and convince them to trial the suggested practice changes while providing them with 'on the ground' technical support. Once the local lead farmers endorsed the program, participant numbers rapidly increased and the program achieved positive outcomes. The early failure of the program was a manifestation of a theory risk. If the program administrator had considered the diffusion of innovations theory (Rogers, 2003), for example, it would have been clear that farmers' practices are more influenced by peers than by experts.

Conclusion

Evaluation of publicly funded programs should provide credible, accurate, timely, and useful information about the current and reasonable future value of a program – with the least use of resources possible. Governments must manage the risks that proposed solutions to problems are insufficient or ineffective. This article suggests that de-risking program design using logic, reasoning, adaptive management, and the sequential purchase of information provides a cost-effective propositional form of evaluation that could be more regularly employed.

Propositional Evaluation is risk-based approach to evaluation. It can be thought of as a dialogue between strategic planners and operational staff, facilitated by evaluators. It is most distinct from other forms of evaluation by focussing on the design phase and prospective or ex-ante evaluation, as well as during delivery. It focuses on managing risk, navigating complexity, and learning within the unknowable, and is explicitly pragmatic, deliberative, and democratic. It invites reasoned discussion about the value of proposed or current courses of action. The necessity and sufficiency of actions and the conditions they bring about in their specific context, rather than abstract theories, are the focus of discussions, debates, and deliberation.

At its core, Propositional Evaluation is quite simple. It is about setting out and testing a claim that if we do these things, we will achieve these conditions. In many ways, it simply supports and extends what many experienced evaluators would think of as traditional good quality evaluation – but it does so by making a fundamental break with the idea of accumulating knowledge about ‘what works’, in favour of managing risks that ‘this works’. It invites a critical reflection on the design of a specific course of action. It is less tolerant of plans with unreasonable assumptions and unjustified outcomes that may sometimes be accepted when a program is offered as a hypothesis or theory to be tested, rather than as an inherently risky but potentially sound, value proposition.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Andrew J Hawkins https://orcid.org/0000-0003-4365-1025

Note

1. We use the terms program, intervention, initiative, or course of action interchangeably in this article.

References

ANU School of Cybernetics. (2022). *Re/defining Leadership in the 21st century: The view from cybernetics [White paper]*. Australian National University & Menzies Foundation. https://cybernetics.anu.edu.au/assets/Redefining_Leadership_in_the_21st_Century-the_view_from_Cybernetics.pdf

Aristotle (1992). *The art of rhetoric. (H. Lawson-Tancred, Trans). (Original work published circa 4th century B.C.E.)*. Penguin Classics.

Bayley, S. (2012). *Evaluation and wicked social problems [Presentation]*: Australasian Evaluation Society.

Bayley, S. (2022). *Why programs fail: The symptoms and causes of underperforming government programs*. Australian Evaluation Society. [Seminar].

Bayley, S., Owen, J., Cummings, R., & Stame, N. (2012). Does performance management have a future? Issues and challenges. In European Evaluation Society Conference, Helsinki, Finland, 3-5 October 201. [Panel discussion].

Bennett, N., & Lemoine, G. J. (2014). What a difference a word makes: Understanding threats to performance in a VUCA world. *Business Horizons*, *57*(3), 311–317. https://doi.org/10.1016/j.bushor.2014.01.001

Bhaskar, R. (2014). *The possibility of naturalism: A philosophical critique of the contemporary human sciences* (4th ed.). Routledge.

Brinkerhoff, D. (1991). *Improving development program performance: Guidelines for managers*. Lynne Rienner Publishers.

Campbell, D. T. (1984). Can we be scientific in applied social science? In R. F. Connor, D. G. Altman, & C. Jackson (Eds.), *Evaluation Studies Review Annual* (Vol. 9, pp. 26–48). Sage Publications.

Canadian Council of Ministers of the Environment. (2021). *Guidance on good practices in climate change risk assessment*. https://ccme.ca/en/res/riskassessmentguidancesecured.pdf

Chalmers, A. (2013). *What is this thing called science?* (4th ed.). University of Queensland Press.

Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and scientific method*. Harcourt Brace and Company.

Datta, L. (1990). *Prospective evaluation methods: The prospective evaluation synthesis*. United Sates General Accounting Office GAO/PEMD-10 (1.10).

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

European Commission, Joint Research CentreRancati, A., & Snowden, D. (2021). *Managing complexity (and chaos) in times of crisis: A field guide for decision makers inspired by the cynefin framework*. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/353

Flood, R. L. (1999). *Rethinking the fifth discipline: Learning within the unknowable* (1st ed.). Routledge. https://doi.org/10.4324/9780203028551

Galef, J. (2021). *The scout mindset: Why some people see things clearly and others don't.* Portfolio/Penguin an Imprint of Penguin Random House LLC.

Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: The RE-AIM framework. *American Journal of Public Health*, *89*(9), 1322–1327. https://doi.org/10.2105/ajph.89.9.1322

Grey, K. (2024). Using formal theory in evaluation – what is it and how is it used? *Evaluation Journal of Australasia*. https://doi.org/10.1177/1035719X241249249

Guba, E. G., & Lincoln, Y. S. (2005). Paradigmatic controversies, contradictions, and emerging confluences. In N. L. Denzin, & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed., pp. 193–215). Sage.

Habermas, J. (1984) *The theory of communicative action: Reason and the rationalization of society* (Vol. 1). Beacon Press. T. McCarthy, Trans.

Hanckel, B., Petticrew, M., Thomas, J., & Green, J. (2021). The use of qualitative comparative analysis (QCA) to address causality in complex systems: A systematic review of research on public health interventions. *BMC Public Health*, *21*(1), 877. https://doi.org/10.1186/s12889-021-10926-2

Hare, J., & Guetterman, T. (2014). Evaluability assessment: Clarifying organizational support and data availability. *Journal of Multidisciplinary Evaluation*, *10*(23), 9–25. https://doi.org/10.56645/jmde.v10i23.395

Hartono, B., Wijaya, D., & Arini, H. (2019). The impact of project risk management maturity on performance: Complexity as a moderating variable. *International Journal of Engineering Business Management*, *11*(10), 184797901985550. https://doi.org/10.1177/1847979019855504

Hawkins, A. J. (2020). Program logic foundations: Putting the logic back into program logic. *Journal of MultiDisciplinary Evaluation*, *16*(37), 38–57. https://doi.org/10.56645/jmde.v16i37.657

Jackson, M. C. (2019). *Critical systems thinking and the management of complexity* (1st ed.). Wiley.

Kurtz, C. F., & Snowden, D. J. (2003). The new dynamics of strategy: Sense-making in a complex and complicated world. *IBM Systems Journal*, *42*(3), 462–483. https://doi.org/10.1147/sj.423.0462

Larson, J. (1980). *Why government programs fail: Improving policy implementation*. Praeger.

Light, P. (2014). *A cascade of failures: Why government fails and how to stop it*. Centre for Effective Public Management, Brookings Institute. https://www.brookings.edu/wp-content/uploads/2016/06/Light_Cascade-of-Failures_Why-Govt-Fails.pdf

Luetjens, J., Mintrom, M., & 't Hart, P. (Eds.), (2019). *Successful public policy: Lessons from Australia and*. ANU Press.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Clarendon Press.

McChrystal, G. S. A., Silverman, D., Collins, T., & Fussell, C. (2015). *Team of teams*. Portfolio Penguin.

Meador, M., Bay, R. C., Anderson, E., Roy, D., Allgood, J. A., & Lewis, J. H. (2023). Using the practical robust implementation and sustainability model (PRISM) to identify and address provider-perceived barriers to optimal statin prescribing and use in community health centers. *Health Promotion Practice*, *24*(4), 776–787. https://doi.org/10.1177/15248399221088592

Mintzberg, H., Ahlstrand, B. W., & Lampel, J. (1998). *Strategy safari: A guided tour through the wilds of strategic management*. Free Press.

Nutt, P. (2002). *Why decisions fail*. Berrett-Koehler Publishers.

Patton, M. Q. (2019). *Blue marble evaluation: Premises & principles*. Blue Marble Evaluation.

Pawson, R. (2013). *The science of evaluation: A realist manifesto*. Sage.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Sage.

Picciotto, R. (2019). Is adversary evaluation worth a second look? *American Journal of Evaluation*, *40*(1), 92–103. https://doi.org/10.1177/1098214018783068

Renger, R. (2015). System evaluation theory (SET): A practical framework for evaluators to meet the challenges of system evaluation. *Evaluation Journal of Australasia*, *15*(4), 16–28. https://doi.org/10.1177/1035719X1501500403

Rogers, E. (2003). *Diffusion of innovations* (5th ed.). Free Press.

Salm, S., Cecon, N., Jenniches, I., Pfaff, H., Scholten, N., Dresen, A., & Krieger, T. (2022). Conducting a prospective evaluation of the development of a complex psycho-oncological care programme (isPO) in Germany. *BMC Health Services Research*, *22*(1), 531. https://doi.org/10.1186/s12913-022-07951-1

Schwandt, T. A. (2015). *Evaluation foundations revisited: Cultivating a life of the mind for practice*. Stanford University Press.

Scriven, M. (1981). *Evaluation thesaurus* (3rd ed.). Edge Press.

Scriven, M. (2008). The concept of a transdiscipline: And of evaluation as a transdiscipline. *Journal of MultiDisciplinary Evaluation*, *5*(10), 65–66. https://doi.org/10.56645/jmde.v5i10.161

Shadish, W., Cook, T., & Leviton, L. (1991). *Foundations of program evaluation*. Sage.

Smith, M. F. (1989). *Evaluability assessment: A practical approach*. Kluwer Academic.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

Toulmin, S. E. (2003). *Return to reason*. Harvard University Press.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. Connell, A. Kubisch, L. Schorr, & C. Weiss (Eds.), *New approaches to evaluating comprehensive community initiatives* (pp. 65–92). The Aspen Roundtable Institute.

Wholey, J. S. (1979). *Evaluation: Promise and performance*. The Urban Institute.

Wholey, J. S. (1987). Evaluability assessment: Developing program theory. *New Directions for Program Evaluation*, *1987*(33), 77–92. https://doi.org/10.1002/ev.1447

Wholey, J. S. (2004). Evaluability assessment. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (2nd ed., pp. 33–62). Jossey-Bass.

Wholey, J. S. (2011). Exploratory evaluation. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (3rd ed., pp. 81–99). Jossey-Bass.

Wong, G., Westhorp, G., Manzano, A., Greenhalgh, J., Jagosh, J., & Greenhalgh, T. (2016). RAMESES II reporting standards for realist evaluations. *BMC Medicine*, *14*(1), 96. https://doi.org/10.1186/s12916-016-0643-1