Lenses and their blind spots: A comparative framework for evaluation paradigms

Andrew J. Hawkins

ARTD Consultants, Sydney, NSW, Australia, andrew.hawkins@artd.com.au

Suggested citation: Hawkins, A. J. (2026). Lenses and their blind spots: A comparative framework for evaluation paradigms. Evaluation. Advance online publication. https://doi.org/10.1177/13563890261449798

Abstract

Evaluation approaches, or paradigms, are best understood as lenses: ways of looking at evaluands that are chosen, not inevitable. Framing them this way presents evaluation traditions as options to be selected rather than positions to be defended, inviting evaluators to ask not 'which paradigm is correct?' but 'which lens best serves this evaluation's context and purpose?' The lens metaphor offers two practical benefits. First, it promotes flexibility, encouraging evaluators to expand their repertoire and guiding them towards the right lens for each situation. Second, it offers a common framework for exploring what evaluation is, making convergence and divergence across paradigms more apparent and improving evaluative discourse and practice. The paper compares six approaches: a-theoretical (or accountability), experimental, realist, constructivist, pragmatic, and systems evaluation, across 10 dimensions, including purpose, causality, validity, and each lens's blind spot. The aim is not to resolve paradigm debates but to structure them usefully.

Introduction

Evaluation is, as Pawson and Tilley (1997) observed, 'a vast and lumbering adolescent' – a discipline that has grown enormously but does not always know where it is going (p.1). This sense of exponential growth is deepened by the academic literature. Once upon a time, the evaluation researcher needed only Campbell and Stanley's (1963) seminal text to look up an appropriate research design and be out into the field. Nowadays, novice evaluators must

consult a veritable almanac of evaluation approaches before they can be confident that they have set out on a useful path. Worse still, adopting one of these approaches often means taking sides in the 'paradigm wars', as positivism and phenomenology, quantitative and qualitative, experimental and constructivist, realist and systems traditions have each developed their own fundamentalist sects. It is difficult terrain. As Shadish et al. (1991) noted, evaluators who try to do everything that Rossi and Cronbach recommend will do little well; some priorities are needed (p.425).

The evaluation literature offers no shortage of attempts to map this bewildering diversity. Classification systems have included generations (Guba and Lincoln, 2005), stages (Shadish et al., 1991), a theory tree (Alkin and Christie, 2004), waves (Vedung, 2010), a garden (Montrosse-Moorhead et al., 2024), a metro map (Lemire, 2024), and a periodic table (Vaca, 2024). Patton (2012) has catalogued over a hundred distinct approaches. Vedung's (2010) wave account traces how evaluation fashions have succeeded one another over time; the present framework takes a different cut, asking not when each paradigm emerged but how they differ structurally across a common set of dimensions.

A recent bibliometric analysis of 30 years of this journal's publications identified eight thematic clusters; evaluative complexity, theory-based evaluation, realist evaluation, dialogue in evaluation, evaluation use, evaluation systems, politics of evaluation, and accountability, and traced a broad conceptual shift over time, in which early emphasis on use and dialogue has given way to growing interest in theory-based approaches, systems, and complexity (Delahais and Thyrard, 2025). Their analysis maps which traditions have been in conversation; the present framework attempts to explain the terms of those conversations, what assumptions about causality, context, validity, and purpose separate the paradigms from one another, and what this means for practitioners choosing between them. Leeuw and Donaldson (2015) identified a related problem: that 'theory' as used in evaluation is fragmented across multiple incompatible meanings, and that the field lacks a structured way to see how different paradigms deploy theory differently rather than treating 'theory-driven' as a single category. What we do not have is a systematic description of how the major paradigms of evaluation agree and disagree with each other across a common set of dimensions. The 10-dimensional comparison offered here addresses that directly.

This paper seeks to lay bare what is in common and what is in dispute among competing paradigms, so that rather than reaching for the hammer and looking for nails in every evaluation, practitioners are supported to choose the lens that will be fit for purpose, extending the idea that the methods tail should not wag the question dog to the level of paradigm selection itself. For the philosophically inclined, this may appear reflective of the Hegelian dialectic, by which progress comes from synthesising opposing arguments into something new and better. This paper does not go that far. Its more modest ambition is to clarify what is in agreement, and what is genuinely in contention, so that when dialectic occurs it is structured rather than circular and may be conducted in a spirit of mutual admiration rather than paradigm warfare.

The word 'paradigm' is used here deliberately rather than the softer 'approach'. What is at stake between the lenses is not merely a preference among methods but competing

worldviews, with distinct commitments on purpose, ontology, epistemology, causality that have not been reconciled. Kuhn's (1962) original use of the term described the coalescence of normal science around a single dominant view, a reading with some historical accuracy in the physical sciences, although even there the emergence of competing paradigms in contemporary physics has unsettled that picture. Social science has had no comparable reckoning. Different disciplines have privileged different paradigms, and multiple paradigms have continued to compete with one another, or more often to talk over and past one another. Evaluation has inherited this condition. The longer ambition of the paper, which reframes paradigms as lenses, is not a denial that the underlying differences are real. It is a move towards methodological and philosophical plurality within a single evaluator, and a prompt to reflect on where the paradigms are commensurable, or at least where each might be useful in different situations.

The article starts with a very brief survey of how the field developed into different paradigms; it summarises the six paradigms and then describes how the dimensions are viewed by each lens. The conclusion promotes paradigmatic flexibility, choosing what is fit for purpose.

A brief survey of the history of evaluation

Evaluation was born of necessity. This brief survey of developments from the 1960s to the early 21st century serves to note the key themes and the coming in and out of fashion of core concepts that structure the framework set out in the body of this paper.

Modern programme evaluation began in the mid-20th century as an applied social science, heavily rooted in experimental methods. As the reach of government expanded into all arenas of life, as societies became more diverse in the values considered appropriate for public discourse, and as appreciation of the science of complexity grew, approaches to evaluation multiplied. In the 1960s, Donald T. Campbell promoted the vision of an 'experimenting society', treating social programmes as experiments to be rigorously tested, often via randomised trials, in pilot form and expanded or dropped based on evidence of effectiveness (Campbell, 1969). Campbell was eager to point out that quantitative designs and impartial measurement of outcomes were important for evaluation to be seen as a scientific tool for learning, not blaming. Around the same time, Michael Scriven argued that evaluation's unique contribution was valuing – determining the merit or worth of programmes, not just measuring outcomes (Scriven, 1991). These early foundations established programme evaluation as a distinct practice focused on evidence and value judgements.

Yet doubts were creeping in. Campbell, Cronbach, Weiss, and Rossi all harboured reservations about the value of experimental evidence from evaluation by the end of their careers. The core doubt was whether the effects of a past intervention provided valid evidence about the value of a similar intervention in another time or place. This was much debated by Campbell and Cronbach, with Campbell falling on the side of internal validity – we need to know that our intervention caused the observed effect – and Cronbach's on

external validity – we need to know whether it would work again somewhere else. Weiss and Rossi took a different approach to this dilemma and focused on understanding the 'theory' that underpins an intervention as the unit of analysis, rather than the programme activities themselves (Weiss, 1995).

At about the same time that dissatisfaction with the utility of experimental data was growing, the constructivist tradition as embodied by Guba and Lincoln (1989) also made its contribution, questioning the utility of scientific approaches and favouring a more dialectical method. Counterbalancing this was the growth of new public management in the 1980s, in which government was expected to operate like a business with performance management as a central concern. Monitoring was joined to evaluation to the extent that 'monitoring and evaluation' were often said in the same breath without pause for what was similar and what was different. Wholey (1979), reflecting the spirit of the times, famously described evaluation as being to government performance what profit is to business performance, suggesting that evaluation must be ubiquitous and critical to understanding the value of government activity.

In the 1990s, realist evaluation provided an update to the internal versus external validity arguments of Campbell and Cronbach, broadly siding with Cronbach's emphasis on external validity (Pawson and Tilley, 1997). This endorsed a scientific approach by focusing on explaining the causal mechanisms by which programmes are effective. So, contra Scriven who held that the science of evaluation was about values, Pawson in his book on the science of evaluation focused on developing knowledge of causal mechanisms operating in context to generate demi-regular outcomes (Pawson, 2013).

In the 2000s, there was a return to the certainty that experimental data could provide, which carried experimental evaluation to the top of the methodological pyramid. This was in some ways a natural outgrowth of the simplicity sought by new public management, carried along by a sudden interest in behaviour by economists and the idea of nudging citizens into making better decisions (Haynes et al., 2012). The scientific aura of randomisation steers clear of stakeholders' values; by emphasising a particular notion of impact evaluation that clinically verifies 'what works', it restored experimentalism as the privileged approach.

At about the same time as this newfound urge for simple answers to complex questions, so arrived systems evaluation with its complex answers to complex questions. As the explosion of things that could and should be evaluated continued, alongside approaches that might provide valid answers to different questions from different perspectives, it remains uncertain whether the capability of evaluators to evaluate, and the capacity of governments to absorb the results, have kept pace.

The six paradigms

This paper proposes that the diversity of evaluation approaches can be usefully organised into six major paradigms. The claim is not that these six exhaust the field, nor that the differences within each paradigm are trivial. The claim is more modest and more practical: that these six represent fundamentally different ways of looking at an evaluand, and that

understanding how they differ – and where they unexpectedly align – provides a basis for making reasoned choices about how to approach any given evaluation.

The accountability or a-theoretical lens is so named for the purpose that motivates its use, though it could have been labelled the 'social impact' or 'outcome measurement' lens had we chosen to focus on the methods it employs. It is not theoretically based (hence why it is described as a-theoretical) like the other lenses and is included to ground the discussion in everyday or non-expert conceptions of evaluation. It treats the evaluand, often a programme, as a kind of promise, for which funding was provided. It is focused on measuring what happened, and then making a comparison to allow for a statement of value. The comparison could be to a standard (such as in a rubric) to what might have otherwise happened (as in a counterfactual), to itself over time, or to the cost of providing the evaluand. This lens is often conflated with the experimental lens in cases where it uses an experimental method to measure outcomes – the difference here is that the goal is not lofty scientific knowledge generation but holding people to account. In practice this lens is often applied with less rigour than its best theorists. The most rigorous articulation of this lens is Scriven's (1991) logic of evaluation, which insists that the standards against which programmes are judged must themselves be grounded in values rather than simply chosen for administrative convenience. For Scriven, producing a merit rating is not a technical exercise but a moral one: it requires identifying whose values count, what criteria flow from those values, and what level of performance against each criterion is sufficient to warrant a claim of merit or worth. His Key Evaluation Checklist (KEC) operationalises this logic as a systematic sequence of checkpoints – from establishing significance and needs, through assessing process and outcomes, to synthesising an overall judgement. Davidson (2005) extended Scriven's framework by making the rubric-building process explicit and accessible, showing how evaluators can combine qualitative and quantitative evidence with value judgements to produce defensible evaluative conclusions. Where Scriven provided the philosophical architecture, Davidson provided the practical methodology. Together they establish that the accountability lens, properly applied, is not merely managerial compliance but values-based judgement: the standard must be justified, not just convenient. The mantra here is 'Value is outcome performance against a justified standard'.

Experimental evaluation has its roots in the philosophical traditions of empiricism and positivism, which emphasise the importance of observable evidence and the scientific method. It aims to establish causal relationships between programmes and their outcomes by comparing results from an intervention group to those of a control group, seeking to control for confounding variables and eliminate alternative explanations. It draws from David Hume's successionist account of causation, in which causal relationships are inferred from the constant conjunction of events rather than from any directly observable causal power. The key commitment is that causal relationships can only be established through the careful manipulation of variables while controlling for potential confounding factors, isolating the effect of a single intervention from its context. The mantra here is 'What works?'

Realist evaluation aims to uncover the underlying mechanisms that generate outcomes in specific contexts. It seeks to explain how and why programmes work by identifying the causal

mechanisms that are triggered in particular contexts, leading to specific outcomes (Pawson and Tilley, 1997). Realist evaluation is chosen here as the most theoretically specific form of theory-driven evaluation, rather than the more ambiguous 'theory-driven' or 'theory of change' terms, to enable sharper comparison with other lenses. Realism is based on the philosophy of Bhaskar and asserts that reality exists independently of our knowledge of it, but that our knowledge is always partial because access to the world is mediated by concepts and theories. Realist evaluation focuses on understanding how programmes work by exploring the interplay between contexts (C), mechanisms (M), and outcomes (O). It recognises that programmes provide resources and opportunities that are interpreted and acted upon by participants in different ways, leading to varied outcomes. The mantra here is 'What works for whom, in what respects, to what extent, and how?'

Constructivist evaluation is a participatory and responsive approach that seeks to create a shared understanding of a programme's value and meaning through dialogue and negotiation among stakeholders. It emphasises the importance of multiple perspectives, local contexts, and the co-construction of knowledge between evaluators and programme participants. Constructivist evaluation has its roots in the interpretivist tradition in the social sciences, drawing on hermeneutics (the theory of interpretation) and phenomenology (the study of lived experience). The central philosophical commitment is that reality is not a fixed external object waiting to be measured but is socially constructed through the interaction between people and their environments. Knowledge is therefore always perspectival, partial, and shaped by the values and position of the knower. Constructivist evaluation also draws on critical theory and the transformative paradigm, which foreground questions of power, voice, and social justice in determining whose knowledge counts. There is some cross-over here with the realism of Bhaskar. The evaluator is not a detached observer but an active facilitator of meaning-making. The mantra here is 'Whose reality counts?'

Pragmatist evaluation is rooted in the philosophical tradition of American pragmatism (Peirce, James, Dewey), which holds that the value of an idea lies in its practical consequences. Michael Quinn Patton's (1978) utilisation-focused evaluation was the defining contribution: it reoriented the field around the principle that evaluations should be judged by their actual use, not their methodological elegance. Joseph Wholey's (1979) framework for the 'sequential purchase of information' reinforced this orientation, arguing that evaluation should invest in further assessment only when the likely usefulness of the added information outweighs the costs of acquiring it. Propositional evaluation (Hawkins, 2020; Hawkins and Bayley, 2025) extends the pragmatist tradition by treating programmes as logical propositions for action rather than theories to be tested, framing evaluation as an ongoing process of inquiry for managing risks that an intervention fails. While the 'what works' mantra may be shared with experimental evaluation, it is more focused on local knowledge and management decisions than the development of theory. The mantra here is 'What makes doing that here and now a good idea?'

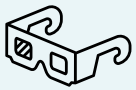
The pragmatic lens gains additional salience in what Schwandt (2019), drawing on Funtowicz and Ravetz (1993), has called 'post-normal' conditions: circumstances in which facts are uncertain, values are in dispute, stakes are high, and scientific reasoning no longer delivers

indisputable conclusions or correct policies. In such conditions, practical reasoning oriented to local knowledge and decision-making, rather than the generation of transferable scientific knowledge, becomes not a methodological preference but an epistemological necessity. Propositional evaluation (Hawkins and Bayley, 2025) can be understood as a systematic response to precisely this condition, asking not whether a programme works in general but whether its assumptions are sound enough to act on now.

Systems evaluation draws from multiple traditions of systems thinking, complexity theory, and cybernetics. Systems thinking focuses on the relationships, interactions, and feedback mechanisms among programme components, stakeholders, and their environments, and how these influence outcomes over time. The approach demands humility: in a complex adaptive system (which is where most policy and programme evaluations exist), no single actor controls the system, knowledge is always incomplete with a short shelf life, causes are multifactorial and distributed across levels, and short-term changes can be unreliable guides to long-term value. Systems evaluation has many iterations both within the evaluation literature (Renger, 2015; Williams and Hummelbrunner, 2010) and outside it (Checkland, 1981; Meadows, 2008). The Cynefin framework provides a helpful diagnostic tool for thinking about the type of system distinguishing between simple, complicated, complex, and chaotic systems, as each requiring fundamentally different responses and approaches to evaluation (Kurtz and Snowden, 2003). The mantra here is 'How do the parts interact?'

These paradigms align on some features and differ on others. Some might be surprised that paradigms sometimes considered in opposition share certain perspectives. For example, realist and experimental evaluation are both focused on the validity of causal claims and on transfactual (or scientific) knowledge but have different attitudes to the unit of analysis and the role of context. Accountability-focused and experimental evaluation are both focused on 'what happened' but have different purpose, the one to report what happened, the other for generating knowledge for future application. These points of convergence and divergence are set out in Table 1 and discussed in detail below. In constructing the table, consistent terminology has been used wherever paradigms share a broadly similar position, so that areas of agreement are immediately visible. What emerges is that no two paradigms align on every dimension, and most share common ground with some lenses while parting company on others.

Table 1. Comparing six evaluation paradigms across ten dimensions.

Dimension	A-theoretical	Experimental	Realist	Constructivist	Pragmatic	Systems
Primary purpose	Accountability to values	Scientific knowledge	Scientific knowledge	Values surfacing	Improvement	Improvement
Temporal orientation	The past (Management)	Endless present (Science)	Endless present (Science)	Endless present (The multiverse)	Here and now (Management)	Constant flux (Management)
Unit of analysis	Programme	Programme	Mechanisms in context (CMOs)	Individual	Programme	System
Ontology of the evaluand	Promise	Hypothesis	Hypothesis	Relationships	Plan	Relationships
Role of context	Distraction	Unwanted intruder	Co-conspirator	King	Co-conspirator	King
Conception of validity	Comparison to a standard or counterfactual	Internal validity of knowledge claims	External validity of knowledge claims	Validity of views: all participants	Logical validity: premises and conclusions	Relational validity: boundaries and actors
Theory of causality	None	Simple (successionist causality)	Realist (generative causality)	Sceptical	Complicated (configurational causality)	Complex (emergent causality)
Role of empirical verification	Core	Core	An ideal	None	Secondary	Secondary
Core evaluation question	How well did it work? (Rubric ratings)	Does it work? (Test causal claims)	How does it work? (Explain causal claim)	Are we transforming society?	Can it work here? Is it working now?	Are relationships healthy?
Major blind spot	Whose values define the standard?	Generalisability/ Replication	Untestable	Politics is real	Short-term	Untestable
Analogous Visual Technology						

The dimensions in detail

Primary purpose

Chelimsky (1997) identified three fundamental purposes of evaluation: accountability, knowledge development, and improvement. These map closely onto the primary purposes served by the six paradigms. The a-theoretical or accountability lens serves, unsurprisingly, the purpose of accountability – providing evidence of what happened and whether public resources were well spent. The experimental and realist lenses serve the purpose of learning – generating knowledge about whether and how interventions work that can be applied in other times and places. Constructivist evaluation is distinct in that its primary purpose is values surfacing – foregrounding the perspectives and experiences of diverse stakeholders. The pragmatic and systems lenses serve the purpose of management or continuous improvement – informing current decisions about what to do here and now.

Making this orientation explicit at the outset can prevent many subsequent disagreements about what constitutes 'good' evaluation. This dimension reveals an important first cleavage. Evaluators who approach their work primarily as accountability practitioners will orient to different methods, evidence standards, and stakeholder relationships than those who approach it primarily as scientists seeking reusable knowledge, or as management consultants seeking to inform immediate action.

Temporal orientation

An evaluator's orientation to the past, the present, or the future reveals much about how they view the knowledge that is and should be generated by evaluation. Three dominant temporal positions can be identified. The accountability lens looks to the past: establishing the facts of what happened in a particular time and space is history, not science, no matter how scientific the methods of data collection and analysis. This orientation infers that past outcomes or past performance are valid predictors of future worthiness for funding. Sometimes the distinction between the past and the present or future is lost on the proponents of experimental design, when the measurement of what happened using an experimental design, or the history of a programme, is conflated with what works, or knowledge about the programme relevant to other times and places and derived from testing scientific hypotheses.

The experimental and realist lenses operate in what might be called an endless present. Experimental evaluation is focused on controlling for context to allow an inference from 'it worked' to 'what works'. But the English language has no specific future tense, so to say 'it works' is inherently ambiguous – it implies the future and refers to the present as though all causally relevant conditions will hold indefinitely. Realist evaluation is similar in its goal of making transfactual statements, with the major difference that 'context' is part of the

equation. These two paradigms clump together on this dimension as both are oriented towards generating knowledge that will be reusable in other similar times and places.

The pragmatic and systems lenses are oriented to a specific place and either a short- or long-term future. They view knowledge generated by evaluation as having a short shelf life, and are focused on informing current decisions about what to do here and now to affect the future. Systems approaches tend to be broad in scope and focused on longer-term emergent properties, while propositional evaluation has the shortest reach of any of the approaches and is the least likely to generate reusable knowledge.

Unit of analysis

This is arguably the most consequential of all the dimensions. The most obvious and traditional focus for the evaluand is 'the program' – a collection of activities designed to achieve some outcome. The accountability, experimental, and pragmatic lenses all take this focus. Realist evaluation does not: it focuses on the purported causal mechanisms that are activated or leveraged by programmes. On a realist account, it is these mechanisms that operate in some contexts or for people in certain circumstances that determine the value of a programme. Thus, the realist unit of analysis is the more 'microscopic' mechanism-in-context configuration.

Constructivist evaluation focuses on individuals and their meaning-making. Systems evaluation takes a broader view, focusing on how the intervention interacts with other existing interventions and with existing structures of the social world. By expanding the scope up or down from the programme itself, realist and systems approaches situate the intervention within existing structures and mechanisms of sociological structure and psychological agency. In both cases, the unit of analysis is defined in terms of interactions between existing forces and those being introduced by an intervention.

Ontology of the evaluand

How the evaluand is conceived shapes everything that follows. The accountability lens treats the evaluand as a promise: one agency sets out a business case or new policy proposition and claims that this approach will deliver certain outcomes. The evaluand is accompanied by a 'program logic' that provides a list of intended outcomes loosely connected by a thread of logic, sometimes just a wish list founded on heroic assumptions.

The experimental and realist lenses treat the evaluand as a hypothesis or theory to be tested. The motivating idea of theory-driven evaluation was to pay attention to the reasons and assumptions, or 'theories of change', that underpin a programme (Weiss, 1995). In science, one develops a hypothesis and then puts it to the test; to a researcher, an evaluand is like this hypothesis, and actual delivery is the test.

The pragmatic lens treats the evaluand as a plan or proposition. This marks a fundamental change in orientation: it is not about measuring outcomes or testing theories but about

ensuring that a proposition or plan is sound. Is the proposition valid and well-grounded? This is a core foundation for propositional evaluation, which shifts from dispassionate measurement towards a tool that assists people responsible for achieving certain outcomes to manage the risks that they do not (Hawkins and Bayley, 2025).

The constructivist and systems paradigms treat the evaluand as a set of relationships. For constructivists, these are relationships between individuals, their contexts, and their meaning-making. For systems evaluators, these are relationships between components, feedback loops, and emergent properties of the broader system.

Role of context

When the focus of evaluation is on accountability and the outcomes generated by a programme, 'context' is often an unhelpful distraction. Funders want to know what happened; they will be impressed by changes in circumstances for participants, whether measured through objective outcomes or ratings against pre-determined rubrics of value. The extent to which outcomes can be attributed to the intervention is usually of secondary importance.

Experimental evaluation is much more seriously engaged with the question of attribution. Its methods are centred on understanding the value of the programme ceteris paribus – all other things being equal. Context is an unwanted intruder and should be removed from the analysis either by experimental or statistical controls. To the realist, the opposite is the case: context is a co-conspirator and is as important as the intervention's causal mechanisms. The pragmatic lens shares this focus on context as a core part of the evaluation, but differs because the unit of analysis remains the programme: context sets the field into which a programme will intervene and is treated in terms of assumptions that are being made about the conditions for a programme.

Systems evaluation takes the veneration of context furthest and sees it as virtually all that is to be explained; the programme is just one feature of a context, and context is king. The constructivist lens agrees: context is king because meaning itself is contextually constituted. Here, we see an interesting cleavage: the accountability and experimental lenses seek to minimise or control context, while the realist, pragmatic, constructivist, and systems lenses embrace it, albeit in different ways and for different reasons. Stame (2010) charts the same cleavage, observing that for counterfactual approaches context is a confounding variable to be controlled, while for goal-free approaches – including realist and systems evaluation – it is constitutive of how programmes produce effects.

Conception of validity

The six approaches each have their own view of validity. Accountability evaluation uses comparison, often combining statistics and stories or using mixed-method evaluation to make a claim about the programme. Experimentalists seek to control for context to measure the independent outcomes of the evaluand and to establish internal validity (Campbell and

Stanley, 1963). Realists are focused on making generalisations from past evaluations to future contexts and are concerned with external validity. Constructivist evaluation grounds validity in the perspectives of all participants.

Systems evaluation has a different relationship to validity: because outcomes are an emergent property of whole systems, they do not lie dormant within one part. If it has any attitude to validity at all, it is relational – based on the validity of the set of relationships under study, the actors included, and the boundaries set to frame the evaluation. Propositional evaluation focuses on the validity of the proposition in logical terms: if the activities were implemented as intended and the assumptions held, would the outcomes follow with near certainty? This is a type of deductive logic based on Aristotle's enthymeme (Hawkins, 2020), accepting that the premises are probabilistic and many assumptions are left unstated.

Theory of causality

Accountability-based approaches are ultimately focused on results rather than understanding causality per se. They may invoke the rhetoric of causality but are uninterested in unpacking it; citizens want improved outcomes, and governments are accountable for delivering them. Constructivist evaluation is openly sceptical of causal claims in the social world.

Causality is central to the experimental and realist lenses, yet there is widespread disagreement on what it means to say a programme 'caused' an outcome. The experimental approach takes a successionist position: causality is inferred from the regular succession of events under controlled conditions. The realist approach takes a generative position: it is mechanisms in context that generate outcomes (Pawson and Tilley, 1997). This feature of realist evaluation is much misunderstood as being about differential effects for different people, when in fact the realist unit of analysis is the invisible causal mechanisms in context.

Systems evaluation sees outcomes as emerging from interrelationships between parts of the system – not predictable from the component parts. The pragmatic lens uses a configurational approach, drawing on the logic of necessary and sufficient conditions. The experimental and realist lenses clump together here as both are deeply invested in causality, while the accountability and constructivist lenses are relatively indifferent to it, and the pragmatic and systems lenses adopt more pragmatic, context-dependent approaches.

The divergence between successionist and generative causality has practical consequences that extend beyond philosophical preference. Because the experimental lens treats context as a confound to be controlled and the realist lens treats context as constitutive of the causal process itself, the two approaches are not studying the same object. Attempts to fuse them, most notably through so-called 'realist RCTs', founder on this distinction. Random allocation is meaningful when the unit of analysis is a programme or treatment; it is incoherent when the unit of analysis is a context– mechanism–outcome configuration, because people cannot be randomly allocated to mechanisms. Mechanisms, on a realist account, are abstract generative structures in the domain of the real, they cannot be observed directly, let alone

assigned. What Bonell et al. (2012) describe as a realist RCT is better understood as a theory-informed positivist design: an improvement on conventional RCT practice, but not realist in any meaningful sense because the unit of analysis remains the intervention rather than the CMO (Hawkins, 2016; Marchal et al., 2013). Nielsen et al. (2024) document the same collision empirically, finding persistent ontological and epistemological tensions when practitioners attempt to operationalise realist evaluation within experimental trial structures. The framework proposed here explains why these tensions resist resolution by design refinement alone: they reflect a genuine incompatibility in what the two lenses take the evaluand to be.

Role of empirical verification

Schwandt (2015), among others, has argued that evaluative claims must be made in the form of sound arguments requiring evaluations that establish facts to support inferences about value. Accountability-based approaches and experimental evaluation are squarely focused on the collection of empirical data and treat empirical verification as core to their approach.

Realists have a more complicated relationship with empirical verification than either the experimentalists or the constructivists. Critical realism locates mechanisms in the domain of the real, with events occurring in the domain of the actual and experiences registered in the domain of the empirical (Bhaskar, 1979). The ontology rules out the direct observation of mechanisms but not empirical engagement with them. The inferential move from the empirical to the real is retroduction, reasoning from observed patterns to the most plausible underlying structures that would produce them.

Realist evaluation operates within this broad tradition, though it reconceives mechanisms more modestly as the reasoning and resources that programmes offer participants, rather than as the deep generative structures of Bhaskar's philosophy of science. Realist interviews (Manzano, 2016) in the realist evaluation tradition provide a key vehicle for gathering evidence on purported mechanisms, eliciting testimony from the reasoning participants bring to a programme, filtered through the cognitive and psychological processes that shape any self-report and interpretation by the evaluator. Where a mechanism has been theorised in advance, interview evidence can serve as a hypothetico-deductive test: the mechanism predicts that participants in the relevant context should report reasoning consistent with it, and the testimony either corroborates the prediction or disconfirms it.

The warrant is strongest in realist evaluation when mechanisms are theorised in advance and tested against fresh evidence, and weakest when they are reconstructed post hoc from outcome patterns. Where realist theory is developed primarily through observation of intra-programme variation, comparing how a programme works across localities or sub-groups, two compounding problems arise. Intra-programme comparison cannot distinguish changes generated by the intervention from those produced by other factors, leaving the attribution question unresolved. It also risks the hasty generalisation fallacy, developing theory that appears to explain observable patterns but is specific to the context in which those patterns were observed, rather than transfactual in the sense a realist philosophy of science requires (Hawkins, 2016). Where this pattern prevails, what is described as empirical verification slides

towards retrospective post hoc explanation, and the Texas sharpshooter metaphor captures the risk: plausible targets painted around clustered outcomes are mistaken for theories that have been tested. Constructivist evaluation, by contrast, steps aside from this debate entirely, treating data not as verification of claims about an independent reality but as co-constructed meaning to be interpreted dialogically.

Systems evaluators and pragmatic evaluators see empirical verification as secondary. As they are not particularly interested in making knowledge claims with anything other than a very short shelf life, empirical verification is about the extent to which premises and conclusions in a value proposition are 'well-grounded' or the extent to which system relationships have been adequately mapped. Here the accountability and experimental lenses clump together as empiricist, while the systems and pragmatic lenses share a more reserved attitude towards data as the primary source of evaluative knowledge for management.

Core evaluation question

The six lenses each have a different core evaluation question. The accountability lens asks: how well did it work? The experimental lens asks: does it work? Facts-based and experimental evaluation often travel together, yet experimental evaluation is only concerned with outcomes for what they reveal about testing hypotheses about the evaluand and whether it will likely work in some other time and place. Realist evaluation asks: how does it work? The overlap in goals between the experimental and realist lenses has generated sustained debate about the possibility of realist experimental designs, a debate this journal has hosted directly (Nielsen et al., 2024). That debate illuminates the framework's explanatory value: the two lenses appear to converge on causality as a shared concern, yet their incompatible theories of what causality is; successionist versus generative, and their incompatible ontologies of the evaluand, programme-as-treatment versus mechanism-in-context, mean that methodological convergence is structurally constrained. Recognising this does not require choosing one lens over the other; it requires being deliberate about which question is being asked. An evaluator who needs to know whether a programme produced a net effect in a past context is using the experimental lens. One who needs to understand which mechanisms, in which contexts, are worth targeting in future interventions is using the realist lens. And one who needs to know whether the programme's assumptions are sound enough to act on now, in this context, with these resources, for these people, is using the pragmatic lens. The choice of lens should follow the question, not the other way around.

The constructivist lens asks: are we transforming society? Systems evaluation asks: are relationships healthy? Its impulse is to map out who is related to what and how the network of relationships is leading, or not, to an emergent outcome of interest. The pragmatic lens asks: can it work here, and is it working now? Its dominant impulse is to check that the design of a programme actually makes sense before or while it is being implemented, identifying what conditions are necessary for a specific plan to be sufficient for bringing about an intended outcome.

Major blind spot

Each lens has a characteristic blind spot. The accountability lens depends on a justified standard, but the question of whose values define that standard is easily overlooked: rubrics can appear objective while embedding the priorities of funders or designers rather than those of programme participants or intended beneficiaries. The experimental lens faces the challenge of generalisability and replication: findings tightly bound to their original context may have limited prospective explanatory value. The realist lens, for all its explanatory ambition, risks generating untestable claims about invisible mechanisms.

The constructivist lens risks underestimating the degree to which political and structural power shapes the outcomes it seeks to surface. The pragmatic lens is, by design, short-sighted; it sacrifices the generation of transferable knowledge for actionable insights in the here and now. And the systems lens, like the realist lens, may generate frameworks that are illuminating but empirically untestable.

These blind spots are features of individual lenses. There is, however, a blind spot that belongs to the framework itself. Schwandt (2019) has proposed that evaluation may be entering 'post-normal' conditions in which the stable paradigmatic categories on which the field has relied are themselves historically situated, shaped by the intellectual and political conditions of their emergence rather than reflecting timeless analytic structures. On this view, the six lenses described here are formations of their era. The framework does not resolve this challenge; it simply makes the current formations explicit enough to be reasoned about and, in time, revised.

Conclusion

This paper has argued that evaluation paradigms are better understood as lenses than as competing truths. Each lens brings certain features of an evaluand into sharp focus and leaves others in shadow. The framework of six lenses and 10 dimensions is not offered as a definitive taxonomy but as a practical tool for structuring conversations that too often stall because the participants are, without realising it, looking through different lenses.

Several patterns emerge from the comparison. No two paradigms align on every dimension, but most share some common ground. The overlaps are partial and uneven: several paradigm pairs share no cell at all, and two dimensions, the core evaluation question and the conception of validity, are unique to each lens by design. The experimental and realist lenses, often cast as rivals, share a commitment to transfactual knowledge and causal explanation but part company on the unit of analysis and the role of context. The accountability and experimental lenses both look backwards at what happened yet differ sharply on whether the goal is to report results or to generate reusable knowledge. The pragmatic and systems lenses both embrace complexity and orient to the future, yet differ in scope: one narrows to the soundness of a specific plan, the other widens to the health of an entire system. These

partial, shifting overlaps suggest that the familiar battle lines of the paradigm wars obscure as much as they reveal.

There is a reflexive quality to this analysis that deserves acknowledgement. How an evaluator sees an evaluand shapes what they think about it, and what they think shapes how they see it. An evaluator trained in experimental methods may instinctively treat context as a confound; one steeped in systems thinking may instinctively treat it as the main event. Neither is wrong, but neither is the whole picture. Making these orientations explicit does not eliminate disagreement, but it does shift disagreement from unreflective habit to reasoned argument. That is a modest but important gain.

The more fundamental aim of this framework is to structure debate about what evaluation is and is not, and where it might go. The proliferation of approaches need not be a source of despair. Much of it reflects genuine agreement: as the comparison shows, paradigms that are routinely cast as opponents share more common ground than their advocates typically acknowledge. Where the differences are real, they are often not methodological but philosophical, concerning what causality is, what counts as valid knowledge, and what evaluation is ultimately for. These irreconcilable differences will not be resolved by better methods. They can, however, be reasoned about, and that is where advances in evaluation theory are most likely to be made. A framework that makes those differences explicit, rather than allowing them to masquerade as technical disputes, is a precondition for the kind of paradigmatic flexibility the field needs.

The practical implication is straightforward. Evaluators should choose lenses as deliberately as they choose methods, matching the lens to the nature of the evaluand, the maturity of the programme, the knowledge that is sought, and the decisions that need to be made. Commissioners and evaluators who negotiate this choice explicitly at the outset are less likely to talk past each other at the end. And reviewers who judge an evaluation's quality should do so with respect to the lens that was agreed, not the lens they happen to prefer.

ORCID iD

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

Funding

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

Declaration of conflicting interests

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

References

Alkin MC and Christie CA (2004) An evaluation theory tree. In: Alkin MC (ed.) Evaluation Roots: Tracing Theorists' Views and Influences. Sage, pp.12–65.

Bhaskar R (1979) The Possibility of Naturalism: A Philosophical Critique of the Contemporary Human Sciences. Harvester Press.

Bonell C, Fletcher A, Morton M, et al. (2012) Realist randomised controlled trials: A new approach to evaluating complex public health interventions. Social Science & Medicine 75(12): 2299–306.

Campbell DT (1969) Reforms as experiments. American Psychologist 24(4): 409–29.

Campbell DT and Stanley JC (1963) Experimental and Quasi-experimental Designs for Research. Houghton Mifflin.

Checkland P (1981) Systems Thinking, Systems Practice. John Wiley and Sons.

Chelimsky E (1997) The coming transformations in evaluation. In: Chelimsky E and Shadish WR (eds) Evaluation for the 21st Century: A Handbook. Sage, pp.1–26.

Davidson EJ (2005) Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation. Sage.

Delahais T and Thyrard A (2025) Conversations in evaluation: Exploring 30 years of debates in our field. Evaluation 32(1): 11–36.

Funtowicz S and Ravetz J (1993) Science for the post-normal age. Futures 25(7): 739–55.

Guba EG and Lincoln YS (1989) Fourth Generation Evaluation. Sage.

Guba EG and Lincoln YS (2005) Paradigmatic controversies, contradictions, and emerging confluences. In: Denzin NK and Lincoln YS (eds) The Sage Handbook of Qualitative Research. 3rd ed. Sage, pp.191– 215.

Hawkins AJ (2016) Realist evaluation and randomised controlled trials for testing program theory in complex social systems. Evaluation 22(3): 270–85.

Hawkins AJ (2020) Program logic foundations: Putting the logic back into program logic. Journal of Multidisciplinary Evaluation 16(37): 38–57.

Hawkins AJ and Bayley S (2025) Managing the risk of program failure: Propositional evaluation as a tool for risk management. Evaluation Journal of Australasia 25(1): 23–44.

Haynes L, Service O, Goldacre B, et al. (2012) Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. UK Cabinet Office Behavioural Insights Team. Available at: https://www.gov.uk/ government/publications/test-learn-adapt-developing-public-policy-with-randomised-controlled-trials (accessed 21 April 2026).

Kuhn TS (1962) The Structure of Scientific Revolutions. University of Chicago Press.

Kurtz CF and Snowden DJ (2003) The new dynamics of strategy: Sense-making in a complex and complicated world. IBM Systems Journal 42(3): 462–83.

Leeuw FL and Donaldson SI (2015) Theory in evaluation: Reducing confusion and encouraging debate. Evaluation 21(4): 467–80.

Lemire S (2024) The evaluation metro map. Journal of Multidisciplinary Evaluation 20(48): 35–40.

Manzano A (2016) The craft of interviewing in realist evaluation. Evaluation 22(3): 342–60.

Marchal B, Westhorp G, Wong G, et al. (2013) Realist RCTs of complex interventions: An oxymoron. Social Science & Medicine 94: 124–8.

Meadows DH (2008) Thinking in Systems: A Primer. Chelsea Green Publishing.

Montrosse-Moorhead B, Schröter D and Becho LW (2024) The garden of evaluation approaches visualization. Journal of Multidisciplinary Evaluation 20(48): 49–58.

Nielsen SB, Jaspers SØ and Lemire S (2024) The curious case of the realist trial: Methodological oxymoron or unicorn? Evaluation 30(1): 120–37.

Patton MQ (1978) Utilization-focused Evaluation. Sage.

Patton MQ (2012) Essentials of Utilization-focused Evaluation. Sage.

Pawson R (2013) The Science of Evaluation: A Realist Manifesto. Sage.

Pawson R and Tilley N (1997) Realistic Evaluation. Sage.

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.

Schwandt TA (2015) Evaluation Foundations Revisited: Cultivating a Life of the Mind for Practice. Stanford University Press.

Schwandt TA (2019) Post-normal evaluation? Evaluation 25(3): 317–29.

Scriven M (1991) Evaluation Thesaurus. 4th ed. Sage.

Shadish WR, Cook TD and Leviton LC (1991) Foundations of Program Evaluation: Theories of Practice. Sage.

Stame N (2010) What doesn't work? Three failures, many answers. Evaluation 16(4): 371–87.

Vaca S (2024) The periodic table of evaluation. Journal of Multidisciplinary Evaluation 20(48): 28–34.

Vedung E (2010) Four waves of evaluation diffusion. Evaluation 16(3): 263–77.

Weiss CH (1995) Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In: Connell J, Kubisch A, Schorr L, et al. (eds) New Approaches to Evaluating Comprehensive Community Initiatives. Aspen Institute, pp.65–92.

Wholey JS (1979) Evaluation: Promise and Performance. The Urban Institute.

Williams B and Hummelbrunner R (2010) Systems Concepts in Action: A Practitioner's Toolkit. Stanford University Press.

Author biography

Andrew J. Hawkins, Chief Evaluator at ARTD Consultants and Vice President of the Australian Evaluation Society, works across paradigms, developing propositional evaluation as an approach to program design and cost effective monitoring and evaluation for managing the risk of failure.