How to make more practical use of complexity-aware evaluation methods

Dr. Tom Aston, Independent

Tom

- 16 years' experience working in evaluation
- Specialises in theory-based & participatory evaluation
- Mostly works in international development
- Recent work for FCDO, World Bank, UK Evaluation Society, European Evaluation Society
- Writes an evaluation blog

Agenda

Time	Description
5 mins	Introduction(s)
15 mins	Introducing complexity
5 mins	Q&A
5 mins	Complexity assessment exercise
10 mins	Group discussion
10 mins	Break
15 mins	Introducing evaluative criteria
5 mins	Q&A
15 mins	Plenary discussion on evaluative criteria
15 mins	Introducing rubrics
5 mins	Q&A
30 mins	Exercise with rubrics
10 mins	Reporting back
15 mins	Introducing bricolage
5 mins	Q&A
5 mins	Discussion
5 mins	Closing

Aims

- Better understand **how to assess complexity and relevant evaluation designs**
- Anchor evaluative choices in **value(able) criteria**
- Increased familiarity with a few **tools to address the challenges of complexity**

Complexity

Complexity: dance between patterns and events

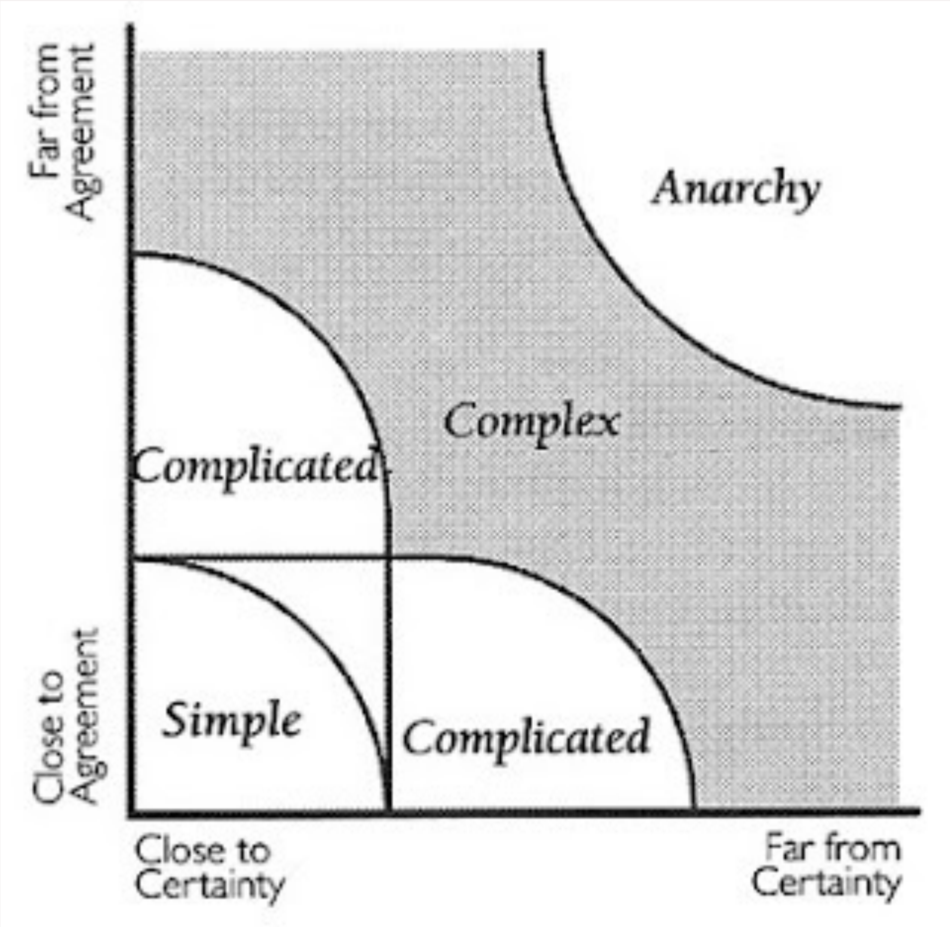
Sensitive to context: One size does not fit all;
Path dependence: Sensitivity to history, initial conditions, & sequence of events;
Systemic: Factors work together interdependently and synergistically rather than independently;
Non-linearity: Outputs and outcomes are not not directly proportional to inputs;
Emergence: Changes are uncertain, and adaptive, but not random;

Episodic: Things are becoming, developing, and changing - change tends to happen episodically, with periods of self-organisation.

Lorenz, 1993; Holland, 1996; Meadows, 2008; Boulton *et al.* 2015; Gates *et al.* 2021

Complexity and evaluative choices

- **Radical** (Patton, 2011, 2015; Mowles, 2014; Jenal and Liesner, 2017) and **pragmatic** (Forss *et al.* 2011; Westhorp, 2012; Bamberger *et al.* 2016; Bicket *et al.* 2020; and Masset *at al.* 2021) positions on complexity



Zimmerman, 2001

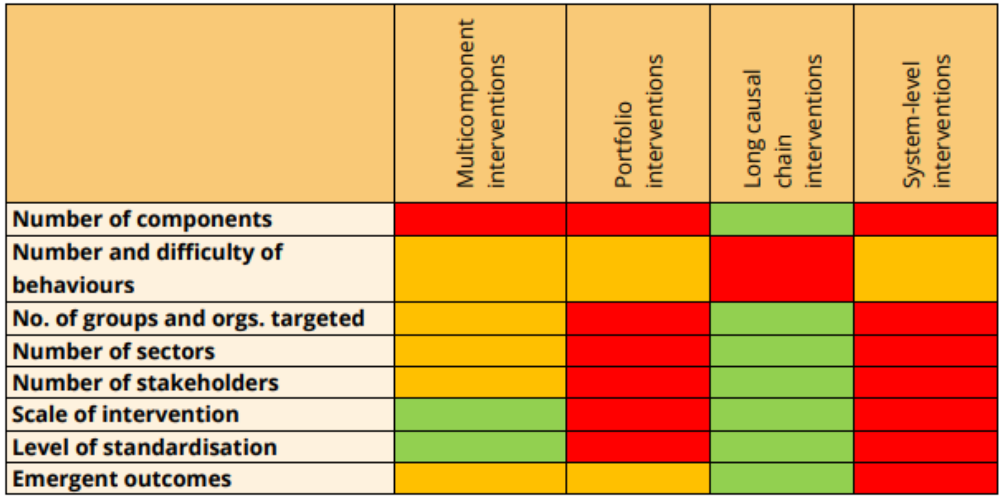
- Taking complexity seriously has implications for: (i) defining evaluation **boundaries**, (ii) scope to assess **interactions between actors and factors**, and; (iii) **emergent outcomes**
- It drives evaluation **questions** and **methodological choices**

Navigating complexity, pragmatically

Project objectives	**Low complexity**: Few and clearly defined objectives **High complexity**: Multiple, broad objectives that are not clearly defined
Nature of the problem the project is addressing	**Low complexity**: Objectives are well understood and there is a high degree of consensus **High complexity**: Objectives not well understood with high level of disagreement
Stability of project design	**Low complexity**: Design is well-tested and stable **High complexity**: Emergent design
Social complexity	**Low complexity**: Implemented within existing social systems and with little conflict **High complexity**: Requires developing new social systems with potential for conflict
Note: The complexity checklist often includes up to 10 dimensions of project complexity, so these are just examples. Each dimension is usually rated on a 5-point scale where 1= low complexity and 5 = high complexity	

Bamberger, 2021

Table 1: Main characteristics of complex interventions



Note: Red cells represent a strong association between intervention type and complexity feature. Orange squares represent a moderate association. Green squares represent low association.

Masset et al. 2021

Causal density: (1) Known technology? (2) Contentiousness in how to respond to the problem? (3) Are there lots of transactions between stakeholders? (4) How much implementer discretion is there? (Woolcock 2013)

See Aston, 2021 for a discussion on complexity checklists

Q&A

Exercise

Consider a project/programme of your own. Give yourself a point for PARTIALLY and two points for YES.

NO	PARTIALLY	YES

- **Knowability:** Are the problems and objectives clearly defined and understood?
- **Certainty:** Has the project/programme met these objectives before here or elsewhere?
- **Consensus:** Do stakeholders agree on appropriate actions to solve the problems?
- **Control:** Is the program relatively independent from context (i.e., isolatable) or do contextual factors significantly affect implementation?
- **Interaction:** Are there many stakeholders and many interactions between them?
- **Stability:** Was program design and implementation relatively stable over time?
- **Discretion:** Is there variation in terms of how actions have been implemented?

Based on Woolcock 2013; Bamberger, 2021; Masset *et al.* 2021

Simple, Complicated, and Complexity-sensitive Methods

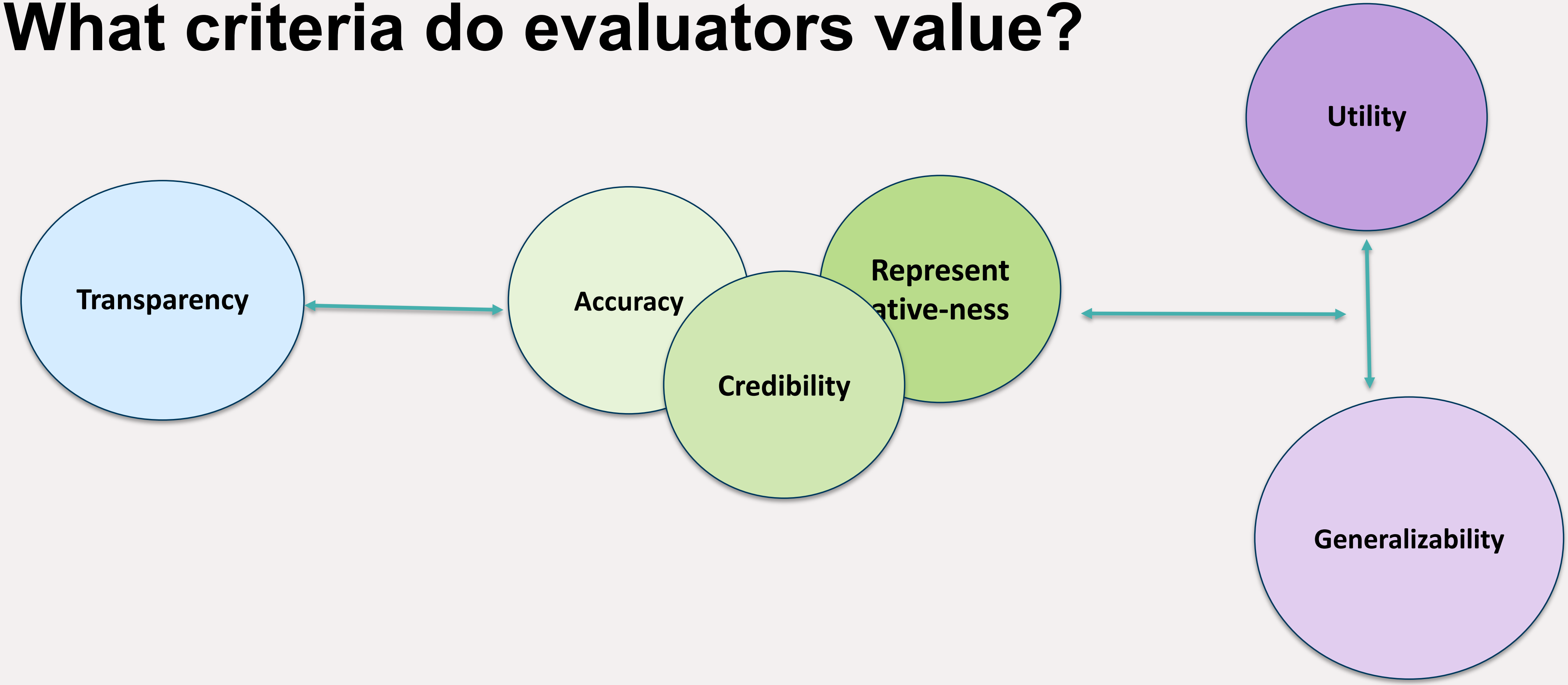
	Simple	Complicated	Complex
Causal Approach	Counterfactual and statistical	Configurational and generative	Generative
Methods	Randomised Control Trials (RCT), Factorial Designs, Adaptive Trials, Propensity Score Matching, Regression Discontinuity, Statistics	Qualitative Comparative Analysis (QCA), Realist Evaluation, Process Tracing, Contribution Analysis	Developmental Evaluation, Action Research, Most Significant Change, Outcome Harvesting, Realist Evaluation, Process Tracing, Contribution Analysis, Causal Loop Diagrams and Systems Dynamics

Adapted from Stern *et al.* 2012, 16-17; Bicket *et al.* 2020: 52 - 58; Masset *et al.* 2021

Break

Evaluative judgement criteria

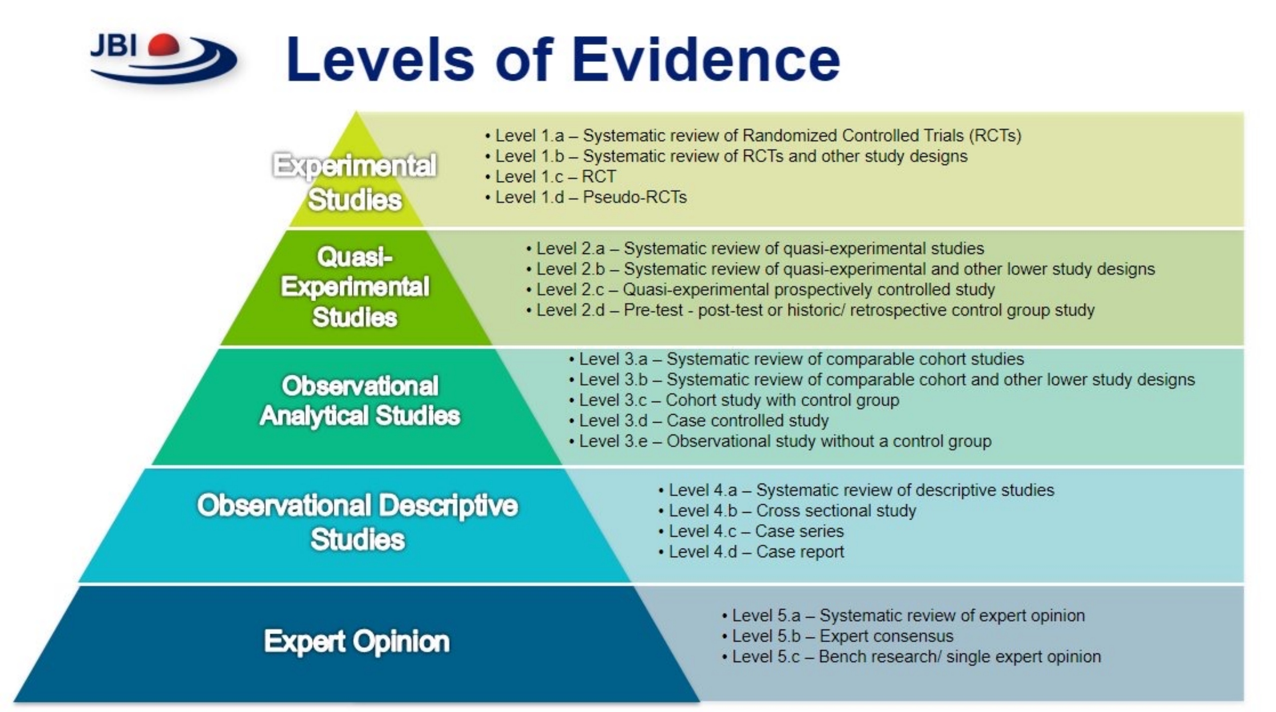
What criteria do evaluators value?



Based on Downes and Gullikson, 2020; see also Aston *et al.* 2021

The fallacy of evidence hierarchies

- Evidence is 'information that has **a bearing on determining the validity of a claim** (*Dictionary of Qualitative Inquiry*).'
- RCTs are not always appropriate, relevant, feasible, or ethical (NICE, 2012; Stern *et al*. 2012).
- **Whether evidence is good or not depends on what it's for – how much it** makes a particular explanation better or worse, or "probative value (Ribeiro, 2019)."
- **"Fragments" or "nuggets" can be powerful evidence** (Pawson, 2008)."



JBI, 2022

Evidence of what, where, and what for?

Evidence standards set the criteria from which we assess evidence claims.

For White (2021), when assessing the **strength of evidence claims** (confidence in evidence claims) we should ask:

1. **Is this evidence relevant? (e.g., question, subject, study, population, location)**
2. **Is it good enough? (i.e., how good is the evidence?)**

Q&A

Discussion

- Which criteria are most important to you and why?
- What are your top 3?
- Is there anything important missing?

The case for rubrics

Rubrics are a form of qualitative scale that denotes a level of performance.

They **explain what the standard *means*** and make clear the **reasoning behind an assessment**.

Rubrics include:

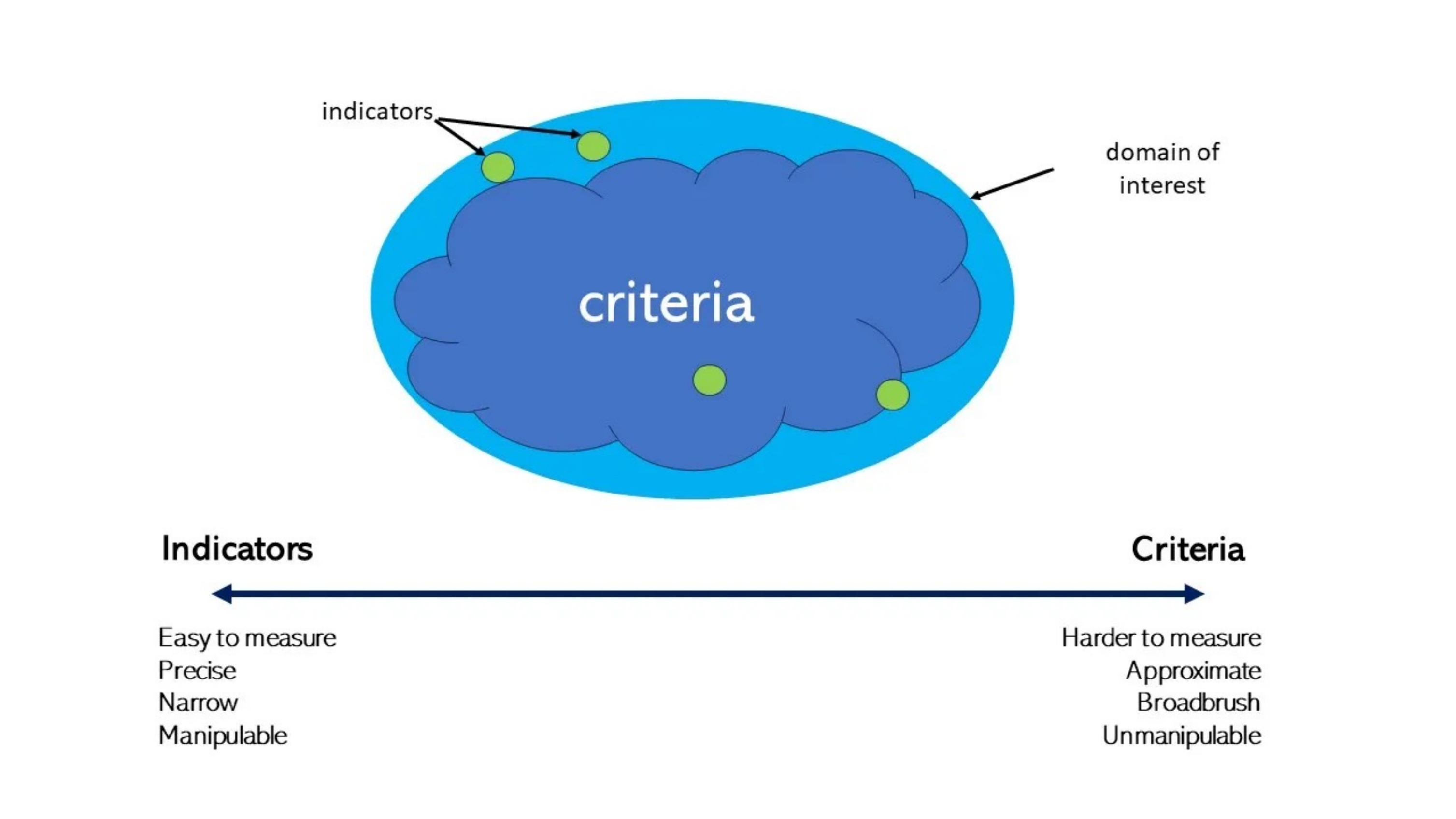
- **Criteria:** Aspects of quality or performance that are of interest, e.g., timeliness
- **Standards:** Levels of performance or quality for each criterion, e.g., poor/adequate/good
- **Descriptors:** Descriptions or examples of what each standard looks like for each criterion in the rubric

Triangulation Rubric

1	2	3	4	5
No evidence corroborates the connection between intervention and outcome. Other evidence contradicts the proposed connection.	A single source of evidence supports the claim connecting intervention and outcome.	Multiple sources of evidence corroborate the connection between intervention and outcome.	Multiple sources of high-quality evidence from different methods corroborate the connection between intervention and outcome.	Multiple sources of evidence across different studies and/or time periods corroborate the connection between intervention and outcome.

Dealing with fuzziness

- Rubrics are appropriate when **boundaries are fuzzy**, and **more culturally flexible, and adaptable**



Oakden, 2011 in Loveridge, 2023

Justifying judgements

One key part underlying assessments in a checklist, scale, or rubric is a **statement which justifies the assessment and final judgement** made:

-**Explain why** a particular grade/score/level was given and;

-**Flag any particular limitations/concerns**

GrammarTop, 2020

Exercise

- **Look at a real case: Developing health technology assessment in Denmark, National Institute for Health and Care Excellence (NICE)**
- Take a few mins to familiarize yourself with the criteria explanations (next slide).
- Each group of 5 looks at 1 criterion to assess an individual case (plausibility, triangulation).
- Groups report back in plenary on their assessment, calling attention to areas of agreement and disagreement.

Using rubrics

Credibility

Triangulation

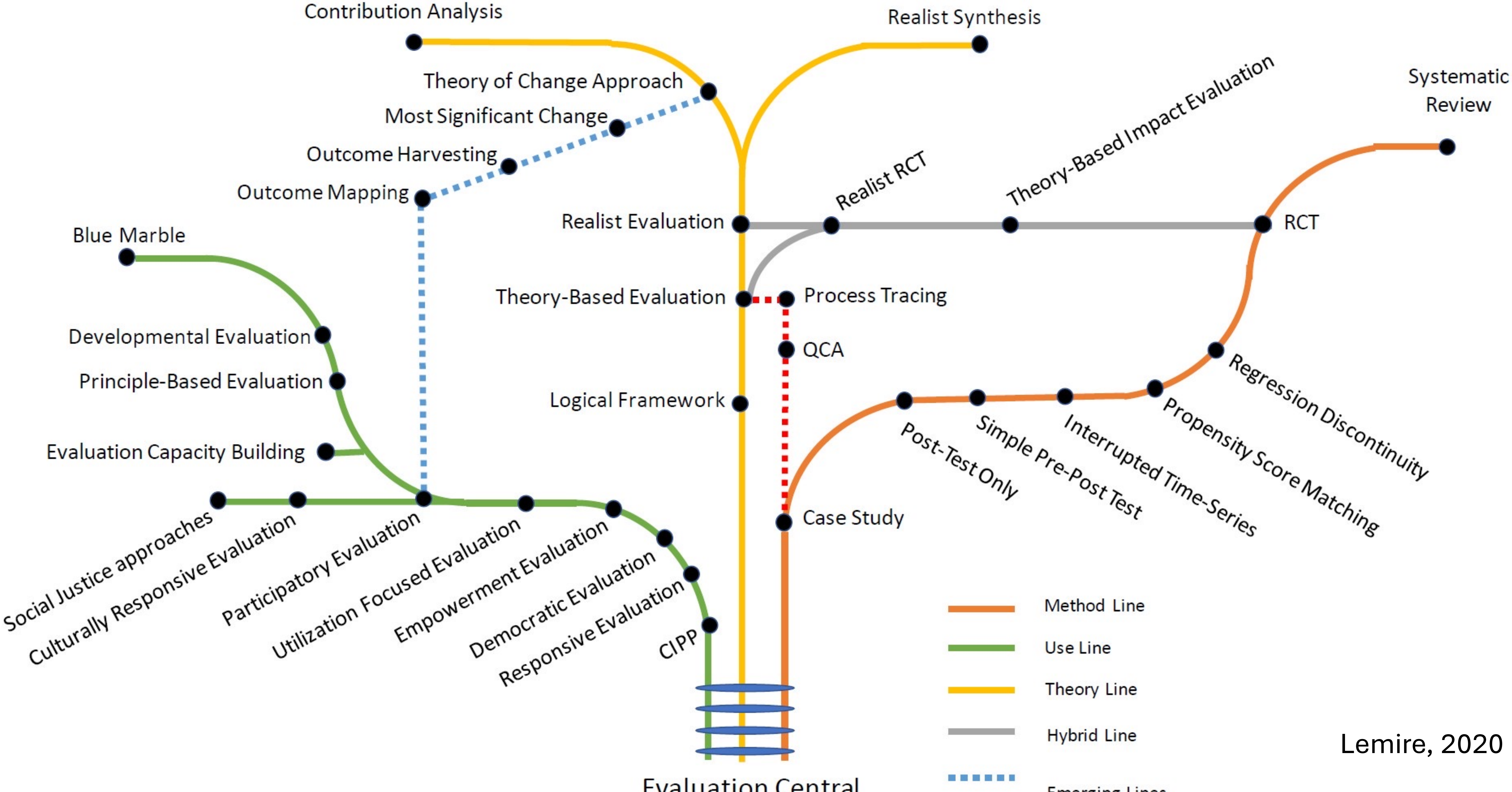
Credibility Rubric

1	2	3	4	5
Unclear, illogical, or contradictory explanation connecting intervention to outcome. Or evidence is found which casts serious doubt on the explanation.	Explanation is clear, logical, and temporally consistent, and suggests a likely association between intervention and outcome.	Relatively convincing account of how evidence connects intervention and outcome. Conclusions drawn mostly follow the data.	Convincing account of how evidence connects intervention and outcome. Conclusions drawn follow the data. Evidence is relatively unique to the explanation and alternative explanations are unlikely.	Highly convincing account, clearly and logically signposting key steps and specific data connecting intervention to outcome. Conclusions drawn unambiguously follow the data. Evidence is highly unique to the explanation and alternative explanations are very unlikely.

Triangulation Rubric

1	2	3	4	5
No evidence corroborates the connection between intervention and outcome. Other evidence contradicts the proposed connection.	A single source of evidence supports the claim connecting intervention and outcome.	Multiple sources of evidence corroborate the connection between intervention and outcome.	Multiple sources of high-quality evidence from different methods corroborate the connection between intervention and outcome.	Multiple sources of evidence across different studies and/or time periods corroborate the connection between intervention and outcome.

Bricolage


Contribution Analysis
Realist Synthesis
Theory of Change Approach
Most Significant Change
Outcome Harvesting
Outcome Mapping
Systematic Review
Theory-Based Impact Evaluation
Realist RCT
Realist Evaluation
RCT
Blue Marble
Theory-Based Evaluation
Process Tracing
Developmental Evaluation
QCA
Principle-Based Evaluation
Logical Framework
Regression Discontinuity
Propensity Score Matching
Interrupted Time-Series
Simple Pre-Post Test
Post-Test Only
Evaluation Capacity Building
Case Study
Social Justice approaches
Culturally Responsive Evaluation
Participatory Evaluation
Utilization Focused Evaluation
Empowerment Evaluation
Democratic Evaluation
Responsive Evaluation
CIPP
Method Line
Use Line
Theory Line
Hybrid Line
Evaluation Central

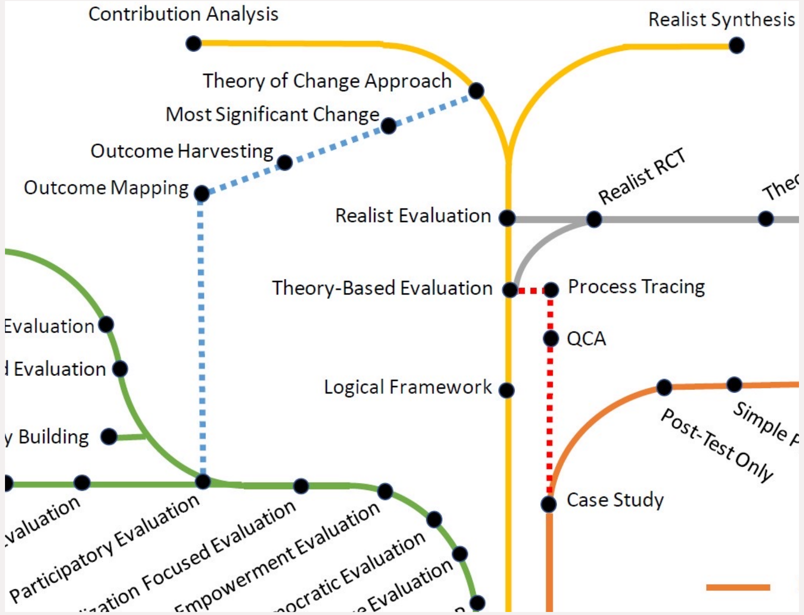
Lemire, 2020

Emerging lines

Case Study

- Qualitative Comparative Analysis
- Process Tracing

… Theory-based evaluation

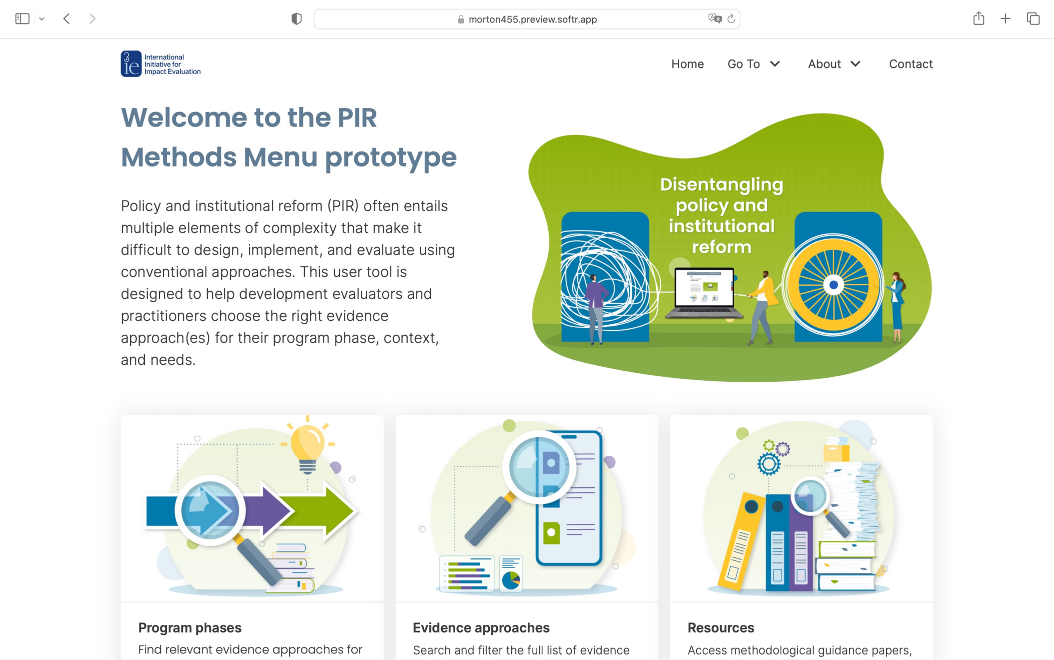


Participatory Evaluation

- Outcome Harvesting
- Outcome Mapping
- Most Significant Change

… Theory of change approach

Choosing *between* methods



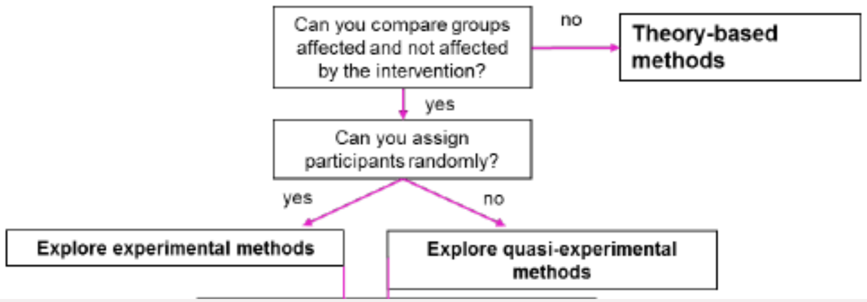
3ie, 2023

- **Either/or options** can be comforting, but misleading
- The common, but insufficient, answer proposed is **“mixed methods”**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	SUMMARY RESULTS - ALL STAGES	RCT (Randomised Contol Trial)	Difference-in-Difference	Statistical Matching	Instrumental Variables (RDD)	Outcome Mapping	Most Significant Change	Soft Systems Modelling	Causal Loop Diagram	PSM (Participatory Systems Mapping)	BBN (Bayesian Belief Networks)	ABM (Agent Based Modelling)	Realist Evaluation	QCA (Qualitative Comparative Analysis)	Process Tracing/ Bayesian Updating	Contribution Analysis
2	Stage 1: Which Method is Best Suited to Answering My Key Evaluation Question(s)?	1	80/100	60/100	1	20/100	40/100	20/100	20/100	20/100	60/100	80/100	20/100	40/100	20/100	20/100
3	Stage 2: Which method is most able to address my other interests?	81/100	60/100	69/100	74/100	47/100	43/100	57/100	59/100	55/100	59/100	62/100	62/100	50/100	59/100	62/100
4	Stage 3: Which Method has the fewest essential methodological requirements that cannot be met by my intervention? (Which method is most feasible to use?)	6	6	6	4	0	0	0	0	0	0	0	0	0	0	0

Befani, 2020

Figure 3.1. Selecting experimental and quasi-experimental methods



HM Treasury, 2020: 47

Increasing method hybridization

Experimental

Theory-based impact evaluation (White, 2009)

Theory-based systematic reviews (White, 2017)

Realist RCTs (Bonnel et al. 2012)

Configurational

QCA (Ragin, 1987)

Realist Evaluation & QCA (Befani et al. 2007)

Generative

Realist Evaluation (Pawson & Tilly, 1997)

Realist Synthesis (Pawson et al. 2004)

Qualitative Impact Protocol (Copestake et al. 2019)

Outcome Evaluation Approach (Belcher et al. 2020)

Contribution Analysis (Mayne, 2001)

Contribution Analysis & Process Tracing (Befani & Mayne, 2014)

Outcome Trajectory Evaluation (Douthwaite et al. 2023)

Process Tracing (Van Evera, 1997)

Contribution Rubrics (Aston, 2019)

Participatory

Outcome Mapping (Carden *et al.* 2001)

Outcome Harvesting (Wilson Grau & Britt, 2012)

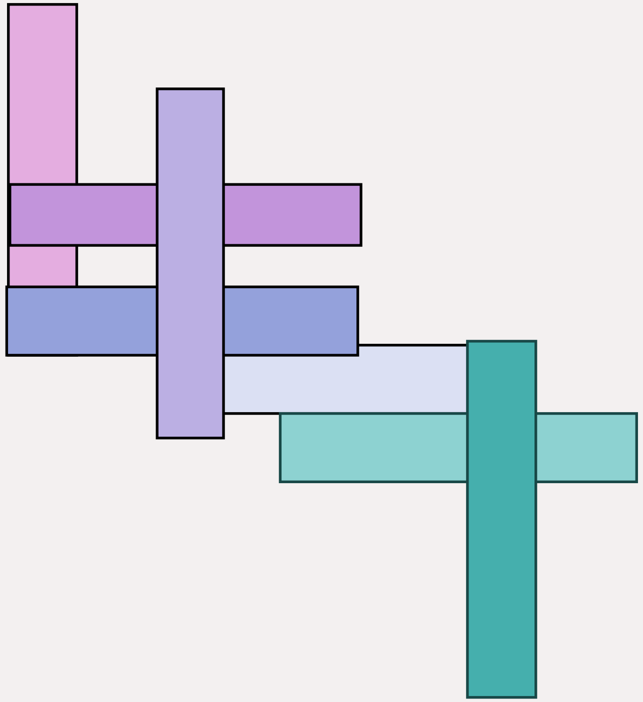
Outcome Evidencing, (Paz-Ybarnegaray & Douthwaite, 2016)

Most Significant Change (Davies & Dart, 2005)

Collaborative Outcomes Reporting (Dart & Roberts, 2014)

Bricolage

- **Bricolage** is a way to combine (or triangulate) the best bits of methods
- Attempt to **reuse a heterogeneous repertoire** of available materials **to solve new problems** (Lévi-Strauss, 1968)
- **Not new**, but gaining prominence in recent years (Patton, 2011; Hargreaves, 2021; Aston & Apgar, 2022)
- **Focus on the parts you most want to strengthen** and tools to strengthen those parts

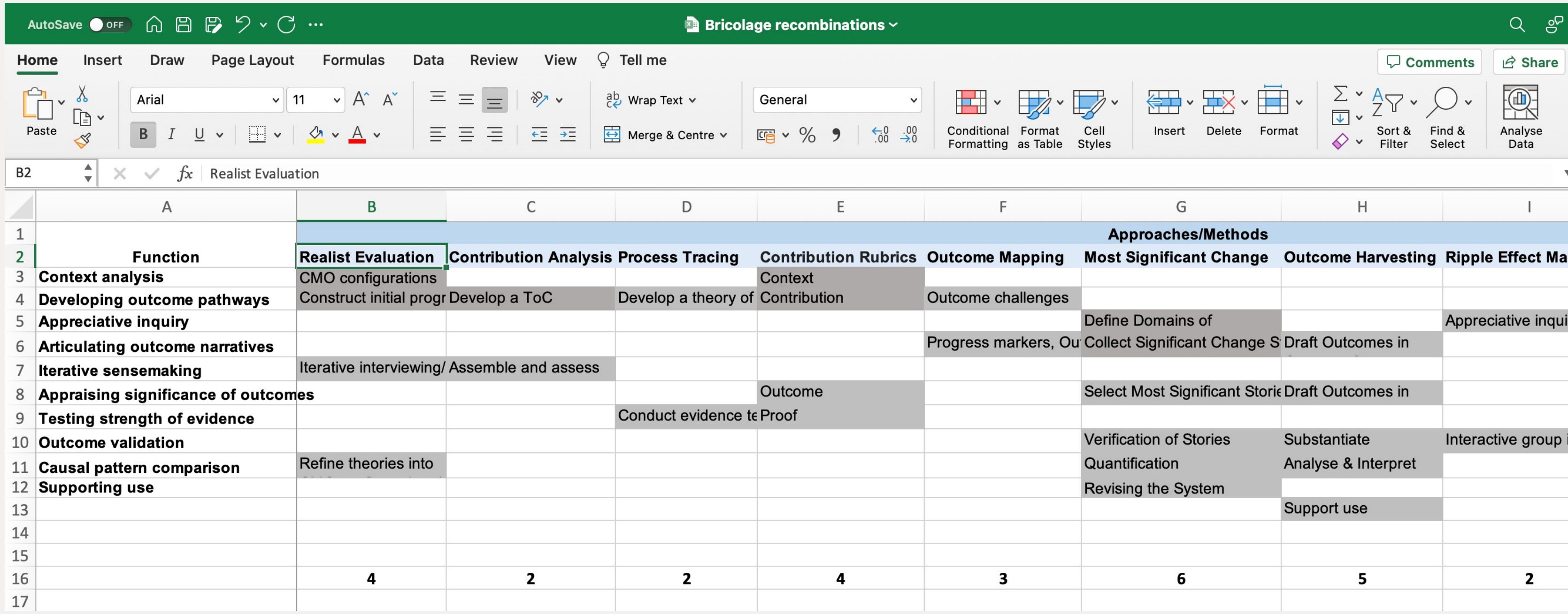


Bricolage criteria

- ***Reasoning***: Critical thinking, alternative explanations and interpretations, and search for outliers
- ***Credibility***: Internal validity, contextually sensitive “probative value” of evidence
- ***Responsiveness***: Reflect local stakeholders’ values and cultural context, sensitive to *their* experiences and definitions of success, and evaluation criteria
- ***Utilization***: Actionable evidence, utilization of evaluation findings
- ***Transferability***: External validity, focused on moderating factors and how the outcomes of an intervention are afforded by the context (Aston *et al.* 2021)

Choosing among options

- Comparable steps, fulfilling **similar functions**
- **Substitutable steps**



	A	B	C	D	E	F	G	H	I
1		Approaches/Methods							
2	Function	Realist Evaluation	Contribution Analysis	Process Tracing	Contribution Rubrics	Outcome Mapping	Most Significant Change	Outcome Harvesting	Ripple Effect Ma
3	Context analysis	CMO configurations			Context				
4	Developing outcome pathways	Construct initial progr	Develop a ToC	Develop a theory of	Contribution	Outcome challenges			
5	Appreciative inquiry						Define Domains of		Appreciative inqui
6	Articulating outcome narratives					Progress markers, Ou	Collect Significant Change S	Draft Outcomes in	
7	Iterative sensemaking	Iterative interviewing/	Assemble and assess						
8	Appraising significance of outcomes				Outcome		Select Most Significant Storie	Draft Outcomes in	
9	Testing strength of evidence			Conduct evidence te	Proof				
10	Outcome validation						Verification of Stories	Substantiate	Interactive group
11	Causal pattern comparison	Refine theories into					Quantification	Analyse & Interpret	
12	Supporting use						Revising the System		
13								Support use	
14									
15									
16		4	2	2	4	3	6	5	2
17									

Finding the right combination: function, not form

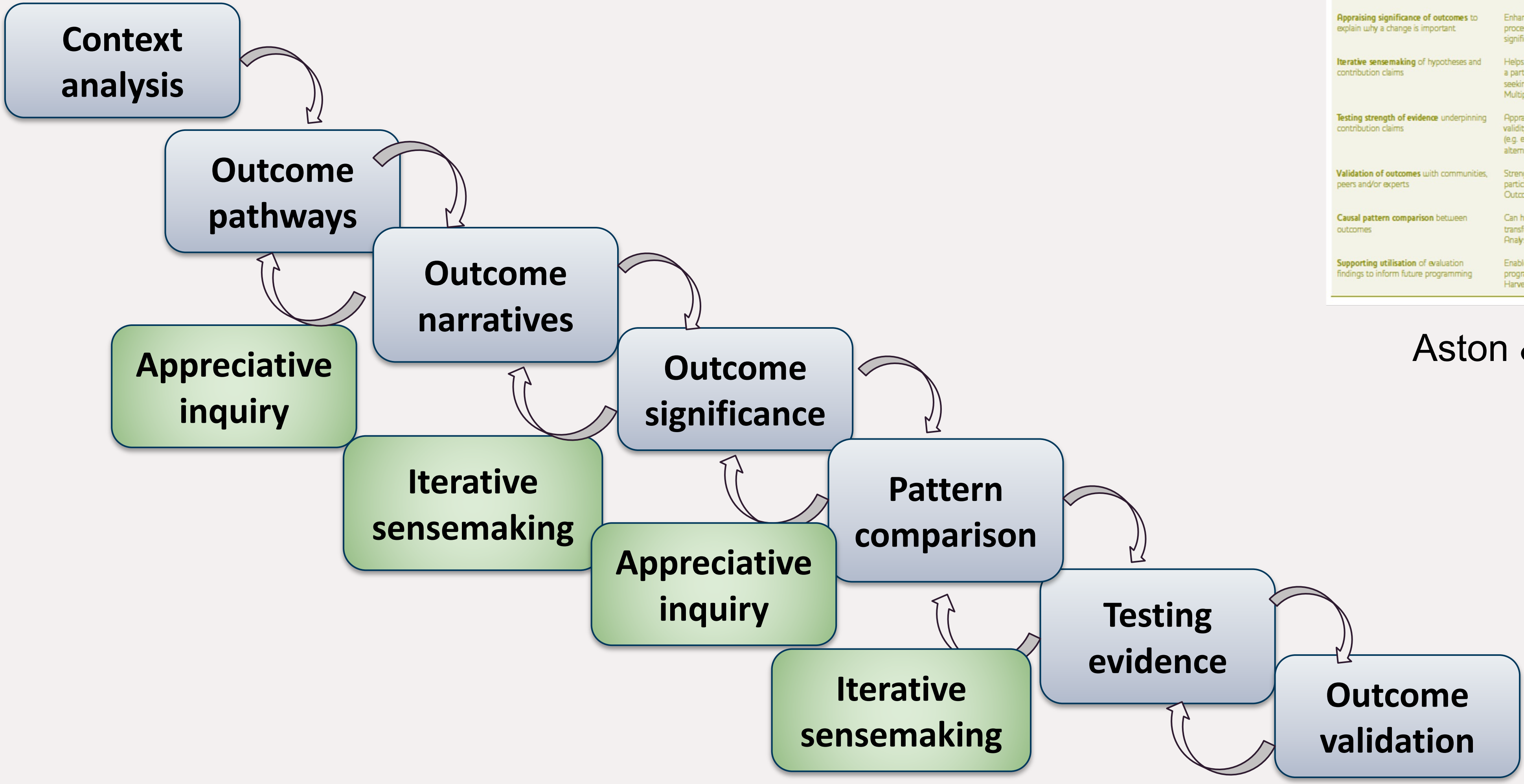


Table 1 **How functions support rigour criteria through methods**

Function	Connection to rigour criteria
Context analysis of potentially contributing factors and conditions	Provides boundaries for causal mechanisms, enhances external validity of evaluative judgements, and supports transferability (e.g. context, mechanism, outcome configurations in Realist Evaluation).
Developing outcome pathways for causal processes	Helps to structure evaluative reasoning, enhances evaluability, and can bolster credibility of findings. If developed through a participatory process, this can contribute to responsiveness and utilisation (e.g. theory of change in Contribution Analysis, or developing causal chains in Process Tracing).
Appreciative inquiry of stakeholders to begin with an understanding of what matters to them, how they envision change, and what they want to learn about	Helps to ensure that evaluation is relevant and representative of stakeholders' experiences, and supports utilisation when orienting evaluation to learn from and about these experiences (e.g. collect significant change stories in Most Significant Change, or social inquiry in Collaborative Outcomes Reporting).
Articulating outcome narratives to explain outcomes and contributions	Enhances evaluative reasoning and the testability of effect patterns, which raises credibility, and can increase responsiveness if these are developed in a participatory way (e.g. draft outcome statements in Outcome Harvesting or progress markers in Outcome Mapping).
Appraising significance of outcomes to explain why a change is important	Enhances evaluative reasoning and improves responsiveness if appraisal is a participatory process (e.g. collecting significant change stories in Most Significant Change or collecting significance data for outcome statements in Outcome Harvesting).
Iterative sensemaking of hypotheses and contribution claims	Helps to strengthen reasoning and enhance credibility and can also be conducted in a participatory way to enhance responsiveness (e.g. assessing a contribution story and seeking additional evidence in Contribution Analysis or revising the conceptual model in Multiple Lines and Levels of Evidence).
Testing strength of evidence underpinning contribution claims	Appraising evidence strength stress-tests evaluative reasoning. It strengthens internal validity and thus increases the credibility of findings and evaluative judgements (e.g. evidence tests and rival hypothesis assessment in Process Tracing; ruling out possible alternative explanations in General Elimination Methodology).
Validation of outcomes with communities, peers and/or experts	Strengthens the credibility of evaluative judgements (internal validity). If conducted in a participatory way, it can contribute to responsiveness (e.g. outcome panel in Collaborative Outcomes Reporting; contribution trial in Process Tracing with Bayesian Updating).
Causal pattern comparison between outcomes	Can help identify trends and outliers, thereby contributing to evaluative reasoning and transferability of findings (e.g. data matrix and truth tables in Qualitative Comparative Analysis; SenseMaker's visualisation of patterns).
Supporting utilisation of evaluation findings to inform future programming	Enables use of evaluation findings and supports transferability through adapting programming and informing new programming (e.g. supports use in Outcome Harvesting).

Aston & Apgar, 2022

Further reading

- Downes and Gullikson, 2020 and Schwandt and Gates, 2021 and Aston, 2022 on **valuing criteria**
- Stern *et al.* 2012; HM Treasury, 2020; Befani, 2020; and 3ie, 2023 on **selecting evaluation methods**
- Aston & Apgar, 2022 on **bricolage** with case-based methods
- Aston *et al.* 2021 and Raimondo, 2023 on **rigour in case-based methods**
- Davidson, 2011; Oakden, 2018; Aston, 2020; King, 2023; and Loveridge, 2023 on the **value of rubrics**