Towards Principled Adequacy for Purpose in Choosing Evaluation Methods

Jonathan Schulte[1*] & Thomas Aston[2**]

[1] LSE Eden Centre for Education Enhancement, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, United Kingdom.

[2] Independent Consultant, UK

* ORCID ID: 0009-0005-0754-0856

** ORCID ID: 0000-0002-4540-8363

Abandoning gold-standard approaches to evaluation methodology renews the challenge of methodological choice and justification. We address this challenge by developing a novel account of methodological assessment we term 'Principled Adequacy for Purpose'. We develop this account by considering recent work centring both the role of questions and values for methodological choice. While we argue that both approaches make important improvements over traditional evidence hierarchies, these frameworks by themselves also face significant limitations. We consider that combining these frameworks, while giving greater consideration to the notion of evaluative purpose, affords better guidance for methodological decision-making in evaluation. For this, we in particularly draw on recent work on Adequacy for Purpose in model evaluation, before combining these approaches under the 'Principled Adequacy for Purpose' umbrella.

There is no gold standard; no universally best method. Gold standard methods are whatever methods will provide (a) the information you need, (b) reliably, (c) from what you can do and from what you can know on the occasion. (Cartwright, 2007: 11)

0. Introduction

A growing literature has highlighted that traditional 'evidence hierarchies' are inappropriate for guiding methodological choices in evaluation (Apgar et al., 2024a; Befani et al., 2014; Blunt, 2015; Cartwright, 2007; Deaton and Cartwright, 2018; Stern et al., 2012). Such hierarchies, ranking methods by their supposed strength of evidence, often tout Randomised Control Trials (RCTs) as the "gold standard" of impact evaluation design, placed atop the evidence pyramid (Crawford et al., 2017; *EEF*, 2016; Farrington et al., 2002). Yet, as Bédécarrats et al. (2019) put it, not all that glitters is gold. RCTs have many limitations, including their difficulties in helping us understand the role of context (Pawson, 2006), their inability to detect heterogenous treatment effects (Blunt, 2015), their potential lack of external validity (Cartwright, 2012), and their limited ability to support project improvement efforts (Stern et al., 2012). Scholars have also pointed out that many real-world contexts fail to support appropriate randomisation (Befani et al., 2014), and, more generally, the often-considerable distance between the idealised assumptions in which proponents of RCTs assess their benefits, and the limitations real-world RCTs encounter in evaluation practice, including lack of blinding, statistical uncertainty, and plausible presence of bias (Deaton and Cartwright, 2018; Krauss, 2021). In short, the claim that RCTs present the universal gold standard is highly questionable.

However, moving beyond evidence hierarchies introduces a new challenge: How to choose from the menu of evaluative methodologies? At least in theory, the benefit of an evidence-hierarchy is that it clearly defines the ideal method to pursue, or, if that is not possible, at least state the confidence with which we should regard evidence from less-than-ideal methods. In contrast, abandoning universal evidence hierarchies requires evaluators to identify fit-for purpose

methodologies and justify them to stakeholders – stakeholders that might still be thinking in 'gold standard' terms.

In this paper, we propose a way to address this challenge by developing the 'Principled Adequacy for Purpose' (PAP) framework. Our argument for this is made in three sections. In section one, we review recent work on evaluation methodological choice: question-first and values-first frameworks. Despite important improvements over evidence hierarchy approaches, we argue that these frameworks are unsatisfactory. In section two, we expand one important reason for this: the neglect of evaluative purpose. Drawing on recent work on adequacy for purpose, we then begin to develop an account of methodological adequacy, which we evaluate in section three. Section four concludes and presents the overall PAP framework.

Before this, it is helpful to briefly define methodology. Following Stirling (2015: 7), we understand methodology to be the broader process through which we determine the merits, capabilities and applicability of specific methods and related techniques and tools. Methodologies are similar to but slightly broader than what Stern *et al.* (2012: 15) call 'designs' which reflect the overarching logic of how the inquiry is conducted (e.g., families of theory-based evaluation or experiments). A method is 'a codified way deliberately to produce knowledge about a focus of interest (e.g. RCT, Process Tracing, or Outcome Harvesting, see Stirling, 2015: 7),' and a tool is an instrument to collect evidence to produce that knowledge (e.g., interview, focus group, survey). As Stern *et al.* (2012: 15) discuss, there are not always perfectly tidy distinctions, as methodology and methods often overlap (e.g., Contribution Analysis or Realist Evaluation). Indeed, as we will argue, selecting fit-for-purpose methods requires a multi-level and iterative process. We will show that making appropriate choices requires consideration of what is valued in the evaluation (axiology), the evaluation's purpose (teleology), the nature of evaluands (ontology), what can be known about those evaluands (epistemology) and practical and ethical design considerations (praxeology) as integral to this broader process (see section 4).

1. Methods Choice as Question-Answer Matching

A first approach, found for example in Stern et al. (2012), Quadrant Conseil's "Impact Tree" (2017), or Befani (2020a), emphasises the importance of choosing methods based on the evaluation questions we want to answer. This perspective also underpins recent work by the Initiative for Impact Evaluation's 'Policy and Institutional Reform Methods Menu' (PIR Methods, n.d.), although the framework adds programme's lifecycle stage as a way to cluster typical research questions by, and subsequently, offers a relatively more fine-grained account of questions. However, as all four approaches consider questions central to methods choice, we jointly term them 'questions-first' approaches.

Concretely, Befani (2020a), tackling what she terms the methods "choice problem", presents the "Refined Design Triangle" (figure 1). In it, she defines methods' appropriateness along three dimensions: (1) their ability to answer the question we want answered; (2) their feasibility in the context of our evaluation project; and (3) methods' "additional abilities" that is, methods' ability to achieve external validity, capture emergent properties, or surface unintended outcomes. Practically, Befani operationalises this process via an Excel tool (Befani, 2020b); inputting the questions to be answered, additional abilities sought, and conditions of the project, the tool assigns the 15 methods included in the Excel tool 'appropriateness' scores between 0 and 100.

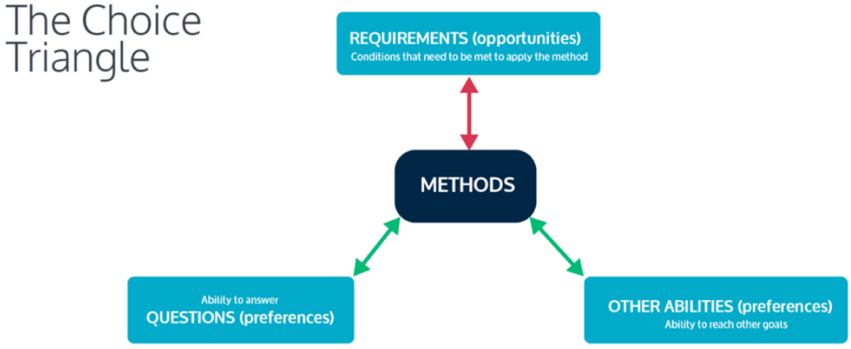

Figure 1: The "Choice Triangle" (Befani, 2020a)

Questions-first approaches thus build on a key insight that Befani attributes to Stern et al. (2012), namely, "that an optimal methodological choice need[s] to align with evaluation questions: what methods are best suited to answer each question?" (Befani, 2020a: 8). Additionally, both Stern et al. (2012) and Befani (2020a) underscore the importance of contextual constraints for applying methods. Across their work, we can identify three dimensions for these constraints: pragmatic constraints – say, the inability to randomise treatments or lack of opportunities to collect certain data; capability constraints, as relevant knowledges and skills may not be available in an evaluation team; and onto-epistemic constraints, concerning the complexity and nature of the programme, acknowledging the limits of what can be known through a particular design.

1.1 Evaluating Questions-first Approaches

Acknowledging the diversity of possible evaluation questions and uneven abilities of different methodologies to answer them highlights an important limitation of evidence hierarchies: RCTs only answer one among many possible evaluation questions. By contrast, questions-first frameworks promise better reasoned methodological choice by ensuring the alignment of questions and the methods. Similarly, Stern et al's and Befani's work highlights the importance of evaluative context. Not all methods can be effectively realised in all contexts or for all evaluands.

Nonetheless, questions-first approaches face important limitations. We can fail to pick the right methodology and methods even when it answers our questions, has the right other abilities and meets our requirements. It is in this sense that the questions-first framework is insufficient to define appropriate methodological choices, and additional factors need to be included.

Consider one of the evaluation question Befani (2020a) discusses: "What was the additional/net change caused by the intervention?" On a standard, counterfactual reading, we are interested in the difference between the actual world, and the counter-factual world of "what-would-have-happened-if-the-program-had-not-been-implemented-but-everything-else-had-been-the-same" (Reichardt, 2022: 160). Befani suggests that RCTs, instrumental variable (IVs) and regression-discontinuity designs (RDDs) are best and equally well suited to answering this question. As Reichardt (2022: 163) argues, however, qualitative methods can equally well answer such questions equally, however. What-if assessments, asking participants "to speculate about how they would have acted if they had not participated in the programme" (Reichardt, 2022: 163) answer the same counterfactual question, providing an estimate of what the world would have been like without the intervention.

Of course, there might be good reasons for preferring RCTs, IVs and RDDs to answer net-change-caused questions. We might question the ability of interviewees to explain what they would have done in the absence of the intervention, raise concerns about the credibility or verifiability of self-assessments, and more generally, doubt the accuracy and unbiasedness of their self-estimates. On such grounds, a "what-if" assessment might well be rejected as a source of credible evidence in situation where an RCT or quasi-experiment would succeed. However, this describes a different kind of methodological failure. We answered the right question but did so in a *sub-optimal way*, failing (implicit) quality criteria which underpin a preferred method. In this sense, it remains unclear how the exact scores Befani (2020a) assigns to different methods are calibrated, or how to weigh up the strengths and limitations listed in the PIR menu. While for example RCTs and RDDs can be considered equally robust in some contexts, their stronger reliance on statistical assumptions (Wing and Cook, 2013) will make them 'weaker' in many others – even relative to the same set of quality criteria. This suggests that a determinant of methodological success remains unacknowledged in questions-first frameworks, specifically, an account of what constitutes 'good answers', beyond their propositional content.

1.2 Methodological Choice as Value Satisfaction

Aston et al. (2022), Apgar et al. (2024a), and Apgar, et al. (2024b) and provide an account for just what this missing dimension is: values. Drawing on foundational work on the role of values in evaluation and research (House and Howe, 1999; Schwandt and Gates, 2021), 'values-first frameworks' aim to explicate stakeholders' values as the basis for methodological choice and evidence assessment in evaluation. This is not a fundamental departure from Befani, who consider that questions to be answered are ultimately grounded in our *preferences* (Befani, 2020a, 2024). However, on Befani's account, these preferences are seemingly taken as given, rather than an expression of what different stakeholders, with the power to decide, value, and debate. In contrast, for 'values-first' scholars, discussion regarding which and whose values count should be the foundation of method and evidence assessment:

> *"At the heart of both valuing and evaluation is criteria – principles or standards that different stakeholders value. Before assessing or rating [evidence], we must first establish what we value – most."* (Apgar et al., 2024a: 101f).

Thus, Apgar et al. (2024a) advise evaluators to facilitate explicit discussions about evidence criteria with stakeholders, identifying what 'good evidence' means to them, based on what they value. They maintain that there are no universal criteria; judgments will depend on local contexts and preferences and are liable to change throughout the evaluation process. Practically, they consider a range of possible values and derived 'quality criteria'. This includes values such as *transparency* about origins of data, *triangulation* across methods and data sources, and *uniqueness,* the ability of evidence to discriminate between alternative explanations for observed outcomes. The authors suggest that evaluation stakeholders work together to formalise their deliberations into evidence rubrics, before rating the produced evidence against these agreed upon standards. This, Apgar et al. (2024a: 110) conclude, "provide[s] a practical architecture for a deliberative process to discuss, debate, an define what success looks like with the main evaluation stakeholders."

1.3. Evaluating the Values-first Approach

Values-first approaches fill an important gap left by questions-first approaches, namely, the need to determine what constitutes 'good answers'. Defining such a notion forces us to deliberate what we value, including our ethical values, which are largely invisible in questions-first approaches[1].

Before proceeding, it is helpful to clarify the values-first approach, pre-empting two possible challenges. The first concerns the admissibility of value judgments in evaluation. This challenge comes in a strong and a weak version. The strong version upholds the 'Value Free Ideal', that is, "that social, ethical and political values should have no influence over the reasoning of scientists" (Douglas, 2009: 1). However, this strong version seems untenable, at the very least in policy evaluation. A Bright (2018) summarises, two lines of argument have been expanded in the literature. First, arguments from *inductive risk* point out that accepting or rejecting a hypothesis – or, in our case, passing evaluative judgement – involves accepting the risk of having arrived at the wrong conclusion. However, such errors bring with them morally significant consequences. Thus, the threshold of 'confident enough' ought to be varied in response to the moral consequences of getting it wrong, meaning that moral value-judgments are central to the job of evaluation. We will return to this argument in greater detail below. Second, arguments from underdetermination point to the gap that inexorably exists between the world and our theorising of it. Because no amount of empirical data can uniquely determine which explanation of framing is 'best', value judgments are involved in determining which approach to pursue and guide myriad decisions along the way. On either argument, values suffuse evaluative inquiry.

While the strong version of the Value Free Ideal should thus be rejected, the weak version of this concern simply holds that not all values are (equally) admissible. For example, Kushner and Stake (2025) agree that values inevitably play a role in evaluation. However, they also maintain that social or political values must not supersede the foundational value of validity in evaluation, that is, must not lead evaluators to make claims outside the 'validity frame' of available data and its reasonable interpretation (Kushner and Stake, 2025: 8). However, values-first approaches are not inconsistent with this view. Instead, they are agnostic: The values-first approach only advocates for us to deliberate and define what we value *in our context,* with values-first authors offering a practical architecture for this deliberation. As part of this deliberation, it might well turn out that we value validity (or some version of it), that we value ethical values more strongly than validity, or that we value both equally.

The second potential challenge to values-first approach is that some values are potentially too abstract to guide practical decision making. For instance, social justice or multi-cultural validity are conceptually dense and may be difficult to translate into concrete evaluative criteria. While this practical challenge can ultimately be addressed through skilful facilitation – ensuring that stakeholders involved in the deliberation process are guided towards sufficiently concrete and operationally useful criteria – we do agree that the values-first framework's practicality can be enhanced. Specifically, we suggest that the values-first approaches, as currently defined can be improved by paying greater attention to notions of evaluative purpose; indeed, the same can be said of questions-first approaches. In the following section, we will therefore outline the case to

[1] Befani's (Befani, 2020b) tool indirectly includes (plausibly) ethical considerations among the 'additional abilities', such as "allowing *all* participants to receive the intervention". However, the ethical dimension of this remains unacknowledged.

re-centre the notion of purpose into our methodological decision-making, defined broadly here as the goal we seek to achieve with our evaluation.

2. Integrating Questions, Values, and Purposes

2.1. Integrating Questions and Purposes

To see how considering purposes enhances questions-first approaches, consider Cartwright's (2012) discussion of the Bangladesh Integrated Nutrition Project (BINP). Despite a rigorous evaluation finding that a near identical programme to have worked in India, the intervention failed in Bangladesh. In Cartwright's reconstruction, this was mainly due to a failure to recognise different social structures across India and Bangladesh, leading to different behavioural responses to the programme and thus impacts (or lack thereof).

Prima facie, we can make sense of this failure from Befani's (2020a) perspective, saying that we asked the wrong question or failed to choose methods with the right additional abilities. As Cartwright emphasises, causal regularities are highly contingent on the system which produces them. As such, we must sharply distinguish evidence for the claim "it works somewhere" from evidence for the claim "it will work for us"(Cartwright, 2012: 976). In other words, "did this work in India?" is a meaningfully different question from "will this work in Bangladesh?" Hence, the purported methodological failure could be considered as having answered the wrong, overly narrow, evaluation question. Following questions-first authors' guidance, we should have picked a method that can answer the right question and provide sufficient external validity.

However, this reconstruction fails to consider the foundational role of evaluative purposes. Judging something to be the right – or wrong – question requires us to identify an aim or goal we want to achieve by answering it. Without a purpose, no question is intrinsically right or wrong. The question "did this intervention work in India?" was answered entirely satisfactorily in the India evaluation. However, it was the wrong question to ask for the purpose of replicate the scheme in Bangladesh. From our perspective, the task of evaluation is not only to answer questions. It is to answer the right question to enable intended uses by producing relevant knowledge for specific evaluative purposes. This means that choices regarding relevant questions should be preceded by a discussion of the purpose of an evaluation. As use-focuses authors consider, we should begin by ask why we are evaluating, and what we want to do with the knowledge we produce (Saunders, 2000). Following this, we can define more appropriate questions.

This proposed phasing for evaluative inquiry is illustrated by Chelimsky's (2006) discussion of how her work for the US government has been shaped by the intertwined purposes of accountability, learning and enhancement. Purposes, combined with high-level policy questions (What are the effectiveness and cost of proposed upgrade options to America's nuclear triad?), were refined into evaluation questions (What is the relative effectiveness of inter-continental ballistic missiles compared to submarine-launched ballistic missiles? What is the cost and value of the proposed upgrades?). It is only at this point that impact evaluation on the level considered by questions-first approaches (what is the net security benefit? How is the deterrence working?) become relevant. Indeed, it is only when considering the expressed purposes of the wider evaluation that it makes sense to prioritise one question over another to guide methodological choice. An analysis aimed as cost efficiency, say, might appear better sustained by a method supporting a counterfactual assessment, whereas questions about the resilience of defensive systems might be better suited to an assessment of the 'mechanism' of

the deterrent – understanding vulnerabilities of the approach through theory-based assessments.

2.2 Adequacy for Purpose

To see how a notion of purpose similarly augments deliberations prescribed by values-first authors, we consider recent work on Adequacy for Purpose, and specifically, its application to the role of values in adequacy (Lusk and Elliott, 2022). Originating in Parker's work on scientific models (2009, 2020), we propose to adapt Parkers' definition from her modelling to our methodology context (while keeping its analytic philosophy prose). Thus, we define methodological adequacy in relation to the use of a methodology and regularity with which it achieves our purpose:

> ADEQUACY_C: M is ADEQUATE_C-FOR-P iff, in C-type instances of use of M, purpose P is very likely to be achieved.

where M is a given methodology or approach, P the evaluation's purpose, and C the specific context of application. In other words, we define a methodology as adequate, given a context and for a purpose, if and only if using it in our contexts is very likely to lead to the achievement our purpose. Methodological choice, on this view, begins by defining our purpose and context.

While our above definition of purpose P as "goal to be achieved" neatly fits into the AFP account (including its role in defining *which questions need answering)*, our context, C, requires refinement. In Parkers' account, C contains and is defined by the range of salient factors we consider affecting the likelihood that using a given approach will enable (or prohibit) us to achieve our purpose. However, her original modelling context yields criteria unhelpful for guiding evaluative methods choices. Hence, the key question is how we can define C to be both true and informative (Alexandrova, 2010: 4).

Notably, Befani's work already allows us to identify two principles of adequacy. M must be able to answer our question; and M must be feasible in our context, that is, it must be possible for us to use our methodology and apply it to our evaluand. A methodology failing on either count appears to be inadequate for purpose, as it cannot tell us what we want to know or cannot be used, prohibiting us from realising its benefits. However, as was argued, these principles are insufficient; we can find feasible methods answering our question, and still fail to achieve our purpose, as our methods might not be ethical, accurate, or transparent enough to realise our purpose.

2.3 Integrating Purposes and Values via Adequacy

Here, the AFP account can help us refine our assessment of relevant values, offering a pragmatic maxim: that 'good' methodology allows us to achieve our primary purpose(s). In turn, the values-first view provides a useful account for determining key aspects of (in)adequacy, by considering values as imposing value-laden criteria which evaluators or stakeholders apply to judge whether an evaluation is good enough. However, this function of defining criteria and deliberating what, in practice, constitutes good evaluative practice points to a subtle role of values that influence concrete decision making and define criteria of good conduct. We call this account an *evaluative praxeology*. This term combines purposeful action (*praxis*) and thought (*logos*), and includes practical knowing (*phrónêsis)* in the service of human betterment (Coghlan and Brydon-Miller, 2014).

While praxeology relates to our broader account of 'what we value' as developed the values-first approaches, it also connects to a goal, and with its focus on appropriate action, must be sensitive to our local context. Two arguments illustrate the need for such a purpose-informed praxeology. The first, offered by Rudner (1953), presents an argument from inductive risk. He considers the decision scientists need to make when determining whether evidence is strong enough to accept or reject a hypothesis. He argues:

> *"Obviously, our decision regarding the evidence and respecting how strong is 'strong enough', is going to be a function of the importance, in the typically ethical sense, of making a mistake in accepting or rejecting the hypothesis" (Rudner, 1953: 2)*

This concern straightforwardly applies to methodological choice as considered here, as our methodology will largely determine the strength of evidence produced. A methodology adequate for assessing whether toxic ingredients are present in a drug ought to produce greater certainty than a methodology 'only' adequate for whether metal buckles are manufactured correctly, as the ethical cost of getting it wrong is much greater in the former than in the latter. Indeed, methods may be (in)adequate for the same phenomenon based on goal we pursue with it: buckles made for seatbelts should require greater methodological care than buckles made for fashion belts (Lusk and Elliott, 2022).

The second argument for the importance of an evaluative praxeology is highlighted by a case-study discussed by Elliott and McKaughan (2014), drawn from Cranor (1995). They consider the California Environmental Protection Agency's choice between two different methods to evaluate the carcinogenic risk of chemicals: one that is more reliable, but slower and more expensive; the other, faster, but less accurate. Ultimately, balancing the value of speed with the value of accuracy and the ethical cost of making a mistake must guide decisions. However, as Elliott and McKaughan point out, a central component of value deliberation in this case concerns *trade-offs* between our values, relative to the goal we want to achieve: how do the risks of wrong decisions compare with the value of quicker decision-making?

Considering these value-laden choices highlights that the elements currently considered by values-first accounts should be augmented. It is not just values, purposes and questions that ought to be deliberated. Good methodological approaches also require judgments of what 'good (enough)' conduct will look like.

2.4 Towards Evaluative Praxeology

In this sense, the values espoused by values-first frameworks offer a potential (non-exhaustive) list *of criteria* influencing whether our methodology is likely to achieve our purpose (s. figure 2). Transparency, for example, is not simply something we value as end in itself; it becomes a requirement that our methods and the evidence produced through them must satisfy, to a certain degree, to satisfy stakeholders and allow us to achieve our intended purpose. However, we can be more specific still by recognising that most discussions on evaluation methodology implicitly construct three distinct aims evaluative aims: an epistemic aim, ethicality, and usability. Indeed, similar distinction can be found across the evaluation literature, including in Scriven (2007), and as implicitly underpinning the branches of Alkin and Christie's 'Theory Tree' (2023).

On this view, the first, central requirement for evaluation to achieve its purpose is epistemic: that it is, broadly speaking, that an evaluation must produce the right kind of insight with the right kind of justification. For ease of reference, we will refer to this epistemic aim as *credibility*,

although we intend to remain agnostic on any metaphysical import. This is as numerous articulation of this epistemic aim have been offered and ferociously debated, with candidates including truth, in all its definitions (Glanzberg, 2023; House, 2014); understanding (Regt, 2017); (sufficient) justification (Rorty, 1998); validity and credibility (Scriven, 2007); and valid argument (Cartwright and Hardie, 2012).

Quality Of Evidence Rubrics For Single Cases (Aston and Apgar, 2023)	Epistemic Values of Evidence-Based Policy (Khosrowi, 2019)
▪ Ethics ▪ Independence ▪ Plausibility ▪ Representation ▪ Transferability ▪ Transparency ▪ Triangulation ▪ Uniqueness	▪ Methodological Rigour[2] ▪ Unbiasedness ▪ Precision

Figure 2. Proposed Values in Evaluation

If we assume that a foundational aim of evaluation is the production of knowledge, we can see that the values in figure 2 are not intrinsic values, but present more concrete 'epistemic values', that is, specific operationalisations that are valuable to the extent that they enhance or ground the credibility of our evaluation (Douglas, 2013; Kuhn, 1979; McMullin, 1982; Steel, 2010). In other words, many of the values proposed are valuable because they are instrumental for achieving our ultimate, epistemic aim (however we define it). This includes values relating to the correctness of our evidence and assurances about the absence of different types of errors, such as representation, triangulation, precision, methodological rigour or unbiasedness; values relating to evidence being collected and 'assembled' in scrutable and logical ways, such as independence and transparency; and values related to the validity or credibility of our *claims relative to other evidence or background assumptions* such as plausibility or uniqueness.

Analogous to the distinction between the epistemic end of evaluation and the lower-level epistemic values and criteria facilitating it, we can distinguish between ethicality as an end of evaluative conduct, and the concrete values and criteria it translates to in practice. Thus, the 'ethics' dimension Aston and Apgar consider (2023) – defined as the requirement that our evidence is produced in ways consistent with our ethical principles – can be broken down further. For example, representation has not just an epistemic dimension, but might also be valued on ethical grounds, representing the value of stakeholders' participation in the process of determining the value and worth of activities; relatedly, transparency has an ethical dimension insofar as it relates to transparently informing stakeholders and participants about the evaluation conduct. Similarly, we may value the absence of harm to participants and their informed consent; or we might value that our findings contribute to equitable outcomes.

The third candidate aim we consider here is usability, that is, "the extent to which the design of an evaluation – both its output and the way it is undertaken – maximizes, facilitates or disables its potential use" (Saunders, 2012: 422). As elaborated by Saunders, while the actual uses our evidence is to some extent unpredictable, certain practices can increase the chances that our

[2] Khosrowi defines methodological rigour in the context of Evidence-Based Policy as 1) a preference for the methods (believed) to be the most reliable; 2) particular care to assessing whether the assumptions of our methods hold; and 3) a general preference for methods requiring the fewest substantive assumptions to begin with.

evidence will be taken up by its intended users in support of our intended uses. Towards this aim of usability, we might value concrete practices such as the timeliness of outputs, the relevance of the evaluation questions to key stakeholders, or the adaptiveness of evaluation designs, that is, their ability to respond to changing and emerging questions throughout the process. Together, these praxeological values are presented in figure 3:

Aim: Credibility	Aim: Ethicality	Aim: Usability
Epistemic Values: ▪ Independence ▪ Methodological Rigour ▪ Plausibility ▪ Precision ▪ Representation ▪ Transferability ▪ Transparency ▪ Triangulation ▪ Unbiasedness ▪ Uniqueness	Ethical Values: ▪ Absence of harm ▪ Informed consent ▪ Equitability ▪ Representation ▪ Transparency	Pragmatic Values: ▪ Timeliness ▪ Relevance ▪ Adaptability

Figure 3: adapted methodological values, sorted into three main dimensions

Distinguishing these different aims underlying our praxeological values has three advantages for determining operational evaluative criteria. Firstly, thus typologised, our praxeology may be better suited to facilitate deliberation of appropriate evidentiary values, emphasising their context-dependent role to act as criteria of our epistemic, ethical and usability aims. Secondly, recognising this distinction means recognising common ground in cases of methodological disagreement. For example, Apgar et al. (2024b) critique experimental methods precisely because they fail to produce valid evidence in complex systems interventions and contexts – remaining on the same epistemic terrain that might motivate defenders of experimental methods to insist on their usage. Indeed, similar to Kushner and Stake (2025) above, we agree that a minimum threshold of validity is necessary for *any* evaluative effort; however, disagreement will arise of what practices best facilitate this epistemic aim in our context, and to what degree a given practice instantiates our validity-criteria. Lastly, a pluralistic conception of evaluative values allows us to explore interdependencies and trade-offs between methodological functions of the kind highlighted by Rudner, Elliott and McKaughan. Insufficiently justified findings might be harmful or not be used in the first place; unethical conduct may limit the degree to which an evaluation is used by policy makers or produces credible data; and methods may increase their usability by becoming available in a timelier fashion, though possibly at the expense of their credibility, with less evidence being collected or being analysed less thoroughly. The relation of *specific* epistemic, ethical, and usability criteria in any one methodological context might be even more complex, as illustrated below (figure 4), especially as we consider that methods vary in their ability to meet different epistemic criteria.

	Epistemic Aim: Will this approach allow us to valid and credible claims?	**Ethical Aim:** Will this approach allow us to meet our ethical criteria?	**Usability Aim** Will this approach be useful (enough)?
Example 1: RCT	If assumptions of an ideal RCT hold (equipoise, control, fidelity, etc.) for a simple intervention context, design closely matches questions to be answered, and assuming that precision, accuracy, reliability, and unbiasedness are valued, the method is likely to produce credible evidence. But from a constructivist perspective, it is unlikely to reflect local stakeholders' values and cultural context or be sensitive to their experiences and definitions of success, and evaluation criteria.	Even under assumptions of equipoise that could justify withholding treatment, possible other ethical desiderata – such as recognising or enabling participants' agency in decision making – are difficult to meet.	Research design – to be 'unbiased' – will need to be highly structured and independent from programme delivery, while demanding high fidelity, which limits the possibility of adaptation based on real-time learning.
Example 2: Outcome Harvesting	If Outcome Harvesting is conducted in a participatory way in a complex intervention context with multiple stakeholders there will likely be high levels of perspectival triangulation and responsiveness to local stakeholders' knowledge, offering more contextually grounded explanations which may contribute to credibility. But from a positivist perspective, certain types of bias, unrepresentative or unreliability might be difficult to rule out, thus limiting credibility in other ways for other audiences with different epistemic values.	Participants' role in producing findings may add additional value by creating reflective space to deliberate work done, valuing participants' agency and creating value for them.	Initial outcome statement collection typically involves programme practitioners, enabling knowledge transfer and providing scope to adapt design to questions arising, increasing the likelihood that findings will influence practice.

Figure 4. Vignettes for comparative adequacy assessment of two methods

As the two vignettes highlight, any real-world assessment of methods relative to values will encounter plenty of "it depends": choices depend on our purpose; on how much emphasise we place on our epistemic, ethical, and usability aims; which values we consider most conducive to these ends; and to what extend we consider that these values are satisfied, in our context, by a given method or approach, including inevitable trade-offs between what we value most.

3. Adequacy For Purpose

Returning to our definition of $ADEQUACY_C$, we can now offer a more substantive definition of our contextual constraints (C). Drawing on praxeological account developed above – and integrating the earlier questions and feasibility dimensions from Befani (2020a) – we arrive at the following definition, along four key dimensions of adequacy in a given context and for a given purpose:

> *Questions & epistemic values*: is the methodology likely to produce valid and credible answers for the question we want to answer? If achieving P in our context requires an assessment of cost-effectiveness, then C's epistemic constraints might demand a large degree of precision and plausible absence of bias. Depending on the values prioritised, an experimental evaluation such as an RCT which require high levels of evaluator independence might be deemed adequate, while a deliberative qualitative approach and participatory methods might be considered inadequate.
>
> *Ethical values*: Is the method likely to produce insights in a way that is consistent with relevant ethical constraints and desiderata? Methods not permitting informed consent and (plausible) absence of harm are likely inadequate for most purposes. However, if P requires considerations of equitability and participant agency, relatively more 'extractive' methods (Cousins and Whitmore, 1998) such RCTs can create problematic power dynamics and negative externalities and may thus be deemed inadequate, while participatory approaches or appreciative interviews might be more adequate.
>
> *Usability values*: is the method likely to produce sufficiently usable evidence? If P requires us to enhance an activity in highly dynamic and changing environments, usability constraints might include preference for methods that are adaptable and produce timely insights. A flexible and often timely approach such as Outcomes Harvesting which includes a specific step for use might be deemed adequate, while more time-intensive theory-based methods such as Process Tracing might be deemed inadequate if timeliness is valued highly (cf. Patton, 2010).
>
> *Feasibility*: is the method likely to be feasible for what we can do and know? This includes the nature of the programme itself – both defining ontology and epistemology, that is, what there is to know and how it can be known (Stern et al., 2012) – and the wider 'ecological' constraints of the evaluation, for example, constraints of data availability or limits to the approaches we can competently apply. A methodology (or particular method) may be appropriate in theory but not in practice if data cannot be collected or analysed in ways that align with a method's standards and protocols. Realist evaluation, given relevant skills on behalf of the research, will be adaptable and thus adequate for many contexts in which data from primary stakeholders can be collected; in turn, RDDs might be inadequate if no suitable discontinuities exist among beneficiaries of an intervention.

In this way, the AFP approach asks us not just to consider our purpose, but also, how our methods will allow us to achieve that purpose in our context, including by considering the

values of our stakeholders and the key questions they hope to be answered. Practically, methodological choice on this account begins by conducting an AFP evaluation: what do we want, and how do we want to get there? Assessing individual methods and their tools and judging whether, across all relevant values, they are jointly adequate or inadequate, we are left with a (possibly empty) set of adequate methods which are justifiable to meet an evaluation's primary purposes. In cases where single choices are preferred, additional criteria can also be added to choose between adequate methods (e.g. choosing the least-costly-but-adequate method).

3.1 Advantages of the AFP Framework

Beyond offering a practical architecture in which to integrate purposes into values-first and questions-first approaches, we consider that an AFP framework has four key advantages.

First, the framework acknowledges that methodological choices are inherently uncertain, defining adequacy via likely achievement of a stated purpose in our context. This means that well-intentioned but poorly executed methods might still end up being inadequate, leaving room for ongoing AFP evaluations and adjustment and refinement. This also means that, along Rudner lines, values will play an important role in determining what likelihood of methodological success is likely enough.

Second, moving away from 'optimal methodological choice', the framework defines a threshold concept of adequacy; several methodologies may well be considered adequate in our context, if they satisfy all our criteria. This could support greater methodological flexibility and exploration, while maintaining a clear lower bound – much but not everything goes. In several respects, this is reconcilable with other approaches. Befani's (2020b) tool has potentially comparable bounds through scores and scale categories, though the question what value defines 'good enough' means would still require additional justification. While not always desirable, following Parker (2020), the AFP framework can accommodate a fully hierarchical account of methods in contexts. We would simply need to define a rank order, such that each methodology is said to fulfil our criteria for appropriateness to a lesser or greater degree or impose an additional measure such as 'cost' by which to rank all adequate methodologies.

Third, the assessment of practices in context allows us to move beyond narrow conceptions of methodology, towards broader 'evidence generating practices.' This can include proposals for multi-method evaluation or "bricolage" by using relevant parts of methods as a mosaic to suit purposes, such as Aston's and Apgar's (2022) discussion of methodological functions. Similarly, purpose satisfaction is also reconcilable with Pawson's (2006) proposal to mine "nuggets of wisdom" from otherwise less credible research by combining or picking elements that jointly satisfy our contextual criteria.

Fourth, the AFP framework allows us to explicate which value influences are legitimate in our methodological context – addressing the concern raised in 1.3, essentially in the same way discussed by values-first approaches. Exploring what adequacy means in context allows us to explore which constraints operate in that context. For example, deliberation with our stakeholders might reveal that usability is valued little, or that certain ethical values are considered as illegitimate influences on methodological choice (Lusk and Elliott, 2022). This contributes to wider discussions on the legitimate role of (non-epistemic) value judgments in scientific practice generally.

3.2 Limitations of the AFP approach

While the adequacy for purpose approach, with its focus on the contextual criteria of goal-achievement, presents an important, integrative alternative to existing frameworks, the approach nonetheless faces three key limitations.

The first limitation is practical: deeply entrenched value preferences will make application challenging. While the AFP approach provides a space for discussion with stakeholders on *why* certain methods should be considered adequate or inadequate, stakeholders perspectives on adequacy may be anchored by evidence hierarchies, with certain methods being seen as inherently superior. Relatedly, RCTs may simply be the only tool adequate for persuading Australia's Treasury Minister Andrew Leigh, author of *Randomistas*, to replicate a programme. This speaks to the political economy of evidence architectures and the need for evaluators to be sensitive to the values, interests, and political calculations of decision-makers (see Dercon, 2025) in their deliberation of how to best produce evidence. In turn, questions of power and influence will inevitably shape local adequacy criteria, raising foundational questions about who gets to determine what matters (cf. Apgar et al., 2024b; Aston et al., 2022).

The second limitation concerns the AFP framework's prescriptive power: the framework is highly flexible. This means that, considering the range of possible purposes and contextual constraints, any methodology might be judged adequate for certain rarified purposes, leaving room for abuse through post hoc rationalisations. Applying an AFP approach in practice thus requires principled action. We need to carefully define what our purposes are in line with espoused values and deliberate what achieving that purpose in our context requires. This, at the very least, imposes a procedural constraint, asking evaluators to provided well-reasoned and informed choices which are clearly rooted in evaluation stakeholders' values and ethical conduct.

The last limitation concerns the framework's completeness: mirroring our discussion of questions-first approaches, the question of the 'right' purpose is exogenous to the framework. Thus, unless we are faced with a well-articulated purpose in the outset of the evaluation – for example in a tendering process – an AFP approach cannot fully guide methodological choice. Cartwright's (2012) discussion illustrates this. In our context C, M_1 (e.g., an RCT) is ADEQUATE$_C$-FOR-P_1 (understanding whether the project worked in India), but inadequate for P_2 (predicting whether the project will work in Bangladesh). Some M_2 (e.g., a cross-country econometric design) is ADEQUATE$_C$-FOR-P_2 but inadequate for P_1[3]. The AFP approach allows us to clearly articulate this inconsistency; however, it cannot answer the prescriptive question, that is, whether to pursue M_1 or M_2. For this, we must settle on either P_1 or P_2 as our purpose. Here, the best way forward appears a return to axiological values: what is it that we and our stakeholders value? And which purpose is more conducive to these values?

4. Principled Adequacy for Purpose

To summarise the argument so far. Considering evidence hierarchies, question-first approaches, values-first approaches, and purpose-first (AFP) approaches, we argued that while all four have some prescriptive power in guiding methodological choices, they all also face important limitations – especially hierarchical approaches, which present the least adaptable approach to methodological choice. In section 3, we further argued that questions, values and

[3] For the sake of argument, we assume P_1 and P_2 to impose inconsistent constraints; in cases where a given M_i is ADEQUATE$_C$ for all candidate Ps, M_i may present a pareto superior choice.

purposes can be integrated under an AFP umbrella – though the question of 'right purpose' remained unanswered. This leads us to propose a synthetic framework intended to draw on the strengths of questions, values, and purpose-focused approaches, which we term '*principled adequacy for purpose'* (PAP). Towards such a framework, we expand previous work by Brown and Dueñas (2020) and Apgar et al. (2024a), incorporating the roles of *teleology* (practical purpose), and *praxeology* (values in action). The resulting iterative map outlining an idealised deliberative process, with key guiding questions, is presented in figure 5.

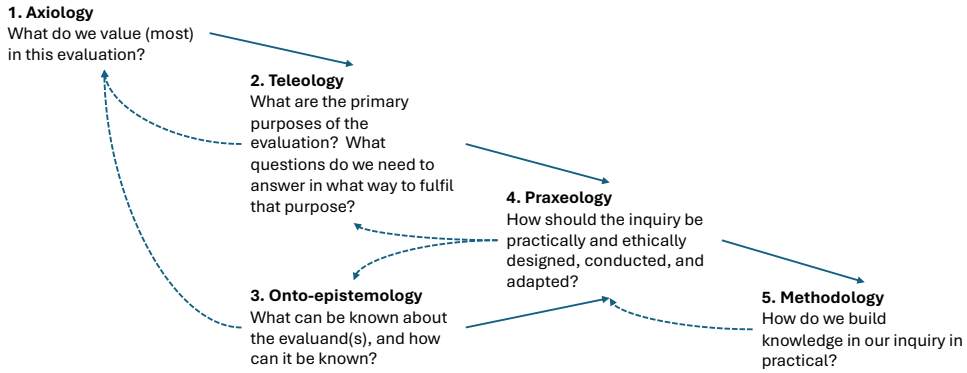

Figure 5: Stylised process map for the iterative deliberation on a Principled Adequacy for Purpose view

Axiology

Our methodological choice process starts with consideration of our axiology (Biedenbach and Jacobsson, 2016; Brown and Dueñas, 2020). This contains a set of the foundational values we hold for the evaluation: the practical, scientific and moral motivations that make us – and our stakeholders – pursue evaluation. Candidate values include validity, credibility, ethicality, usability, or even more broadly, integrity, social justice, well-being, democracy. These deliberations define and ground the foundational motivations of the inquiry. As argued in 2.4, we consider that three candidate goals – the epistemic, ethical and usability function of evaluation – will likely play a considerable role in our axiological foundations.

Teleology

Next, we consider teleology: the account of the goal we want to pursue with our evaluation. Deliberation at this level includes several elements: how our purpose fits with our axiological values; what questions we consider this purpose to raise; the possible operationalisations of axiological values in more concrete categories (from "ethics" to "representation"); deliberation of what potential trade-offs between values, aims and purposes we are willing to accept, and what minimum standards are required. How quick is quick enough to achieve our purpose? What level of risk are we willing to take? Candidate values at this level might include, following Chelimsky, accountability, learning, or enhancement, and, derivatively, the value of speed relative to accuracy. Subsequently, our notion of teleology includes Befani's 'questions' and 'other abilities' dimensions but remains ultimately wider and explicitly grounded in values.

Onto-Epistemology

In our approach, our onto-epistemology contains the set of beliefs about what exists and how it can be known. We use the term onto-epistemic, as popularised in feminist and decolonial scholarship, to highlight the close link between being and knowing (Gatt, 2023). We also note that the onto-epistemic framing of our project is separate from the stream of values-led considerations, reflecting the importance of facts external to the evaluation to constrain our

otherwise values-led reasoning. Deliberations at this stage concern the (best) understanding of the entities under evaluation, for example as persons, programs, ideas, policies, products, systems, performances, that is, an ontology; and foundational epistemic criteria like logical consistency or empirical adequacy that set out the basic constraints of how we can know about these entities. This dimension closely resembles the 'programme attributes' in Stern et al's (2012) work, as argued above.

Praxeology

Next, we consider our praxeology, the account of good purposeful action and conduct in a particular evaluative context (Coghlan and Brydon-Miller, 2014). Deliberation at this stage brings together both the teleological question (what must our evaluation be like to achieve our purpose?) and the constraints imposed by the onto-epistemic account. Thus, it involves developing the operational criteria and standards, identifying and navigating epistemic, usability and ethical trade-offs, and navigating tensions between our values and the 'worlds' onto-epistemic impositions. At the praxeological level, these impositions additionally take on a pragmatic character: based on the nature of the evaluand and our requirements of knowing it, we must consider what we can know on the occasion. Limits on the data we collect, and methodological proficiencies will act as important constraints on our conduct and its ability to be 'good enough' to achieve our purpose.

Methodological Choice

Choosing preferred methods then, should be the *culmination* of the aforementioned systematic process through which we determine the merits, capabilities, and applicability of underlying methods (codified ways to produce knowledge) and related techniques and tools (i.e., instruments) to build knowledge through our inquiry in a particular evaluative context regarding a focus of interest (see Stirling, 2015: 7).

6. Conclusion

To choose and justify methodology appropriately, we need effective frameworks to guide choice and enable us to justify our choices to other. We show that current approaches – centering the questions and values as primary guides to decision-making – contain important ingredients for such an account, but ultimately, remain insufficient to guide choices. These approaches also neglect the role of evaluative purpose. To overcome these limitations, we develop an integrative account, termed 'Principled Adequacy for Purpose' that we argue can better guide methodological choice by recognising the context-dependent role of praxeological values, questions and feasibility constraints to act as criteria our methodological choices must meet *in the right way* to enable us to achieve our purposes.

References

Alexandrova A (2010) Adequacy for Purpose: The Best Deal a Model Can Get. *The Modern Schoolman* 87(3/4): 295–301.

Alkin MC and Christie CA (eds) (2023) *Evaluation Roots, Third Edition: Theory Influencing Practice*. 1st edition. New York ; London: Guilford Press.

Apgar M, Aston T, Snijder M, et al. (2024a) Raising the Bar: Improving How to Assess Evidence Quality in Evaluating Systems-Change Efforts. *The Foundation Review* 16(2).

Apgar M, Bradburn H, Rohrbach L, et al. (2024b) Rethinking rigour to embrace complexity in peacebuilding evaluation. *Evaluation* 30(3): 408–433.

Aston T and Apgar M (2022) *The Art and Craft of Bricolage in Evaluation*. Number 24, Practice Paper, 14 October. The Institute of Development Studies and Partner Organisations. Available at: https://opendocs.ids.ac.uk/articles/report/The_Art_and_Craft_of_Bricolage_in_Evaluation/26433694/1 (accessed 15 May 2025).

Aston T and Apgar M (2023) *Quality of Evidence Rubrics for Single Cases*.

Aston T, Roche C, Schaaf M, et al. (2022) Monitoring and evaluation for thinking and working politically. *Evaluation* 28(1): 36–57.

Bédécarrats F, Guérin I and Roubaud F (2019) All that Glitters is not Gold. The Political Economy of Randomized Evaluations in Development. *Development and Change* 50(3). International Institute of Social Studies: 735–762.

Befani B (2020a) *Choosing Appropriate Evaluation Methods - A Tool for Assessment & Selection (Version 2)*. CECAN. Available at: https://www.cecan.ac.uk/wp-content/uploads/2020/11/Final_Choosing-Appropriate-Evaluation-Methods-1.pdf (accessed 15 May 2025).

Befani B (2020b) Choosing Appropriate Evaluation Methods Tool (Version 2.1). CECAN.

Befani B (2024) Combining qualitative and small-n methodologies in impact evaluation design (Dr Barbara Befani). Available at: https://www.youtube.com/watch?v=evcFBJ34Gpw (accessed 19 May 2025).

Befani B, Barnett C and Stern E (2014) Introduction – Rethinking Impact Evaluation for Development. *IDS Bulletin* 45(6): 1–5.

Biedenbach T and Jacobsson M (2016) The Open Secret of Values: The Roles of Values and Axiology in Project Research. *Project Management Journal* 47(3). SAGE Publications Inc: 139–155.

Blunt CJ (2015) *HIERARCHIES OF EVIDENCE IN EVIDENCE-BASED MEDICINE*. London School of Economics and Political Science. Available at: http://etheses.lse.ac.uk/id/eprint/3284 (accessed 15 May 2025).

Bright LK (2018) Du Bois' democratic defence of the value free ideal. *Synthese* 195(5): 2227–2245.

Brown MEL and Dueñas AN (2020) A Medical Science Educator's Guide to Selecting a Research Paradigm: Building a Basis for Better Research. *Medical Science Educator* 30(1): 545–553.

Cartwright N (2007) Are RCTs the Gold Standard? *BioSocieties* 2(1): 11–20.

Cartwright N (2012) Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps. *Philosophy of Science* 79(5): 973–989.

Cartwright N and Hardie J (2012) The Theory That Backs Up What We Say. In: Cartwright N and Hardie J (eds) *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford University Press, p. 0. Available at: https://doi.org/10.1093/acprof:osobl/9780199841608.003.0002 (accessed 15 May 2025).

Chelimsky E (2006) The Purposes of Evaluation in a Democratic Society. In: *The SAGE Handbook of Evaluation*. 1 Oliver's Yard, 55 City Road, London England EC1Y 1SP United Kingdom: SAGE Publications Ltd, pp. 34–55. Available at: https://methods.sagepub.com/book/the-sage-handbook-of-evaluation/n1.xml (accessed 15 May 2025).

Coghlan D and Brydon-Miller M (2014) Praxeology. In: *The SAGE Encyclopedia of Action Research*. SAGE Publications Ltd, pp. 651–653. Available at: https://methods.sagepub.com/ency/edvol/encyclopedia-of-action-research/chpt/praxeology (accessed 25 May 2025).

Cousins JB and Whitmore E (1998) Framing participatory evaluation. *New Directions for Evaluation* 1998(80): 5–23.

Cranor CF (1995) The Social Benefits of Expedited Risk Assessments. *Risk Analysis* 15(3): 353–358.

Crawford C, Dytham S and Naylor R (2017) *The evaluation of the impact of outreach: proposed standards of evaluation practice and associated guidance*. *Office for Fair Access: Bristol, UK*. Report, June. Bristol, UK: Office for Fair Access. Available at: https://www.offa.org.uk/egp/improving-evaluation-outreach/ (accessed 15 May 2025).

Deaton A and Cartwright N (2018) Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* 210. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue: 2–21.

Dercon S (2025) Best buys meet political realities: The political economy of education research. In: *VoxDev*. Available at: https://voxdev.org/topic/education/best-buys-meet-political-realities-political-economy-education-research (accessed 19 May 2025).

Douglas H (2013) The Value of Cognitive Values. *Philosophy of Science* 80(5). [The University of Chicago Press, Philosophy of Science Association]: 796–806.

Douglas HE (2009) *Science, Policy, and the Value-Free Ideal*. PIttsburgh, UNITED STATES: University of Pittsburgh Press. Available at:

http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=2038867 (accessed 25 May 2025).

EEF (2016) EEF Blog: Do EEF trials meet the new 'gold standard'? Available at: https://educationendowmentfoundation.org.uk/news/do-eef-trials-meet-the-new-gold-standard (accessed 15 May 2025).

Elliott KC and McKaughan DJ (2014) Nonepistemic Values and the Multiple Goals of Science. *Philosophy of Science* 81(1): 1–21.

Farrington DP, Gottfredson DC, Sherman LW, et al. (2002) The Maryland Scientific Methods Scale. In: *Evidence-Based Crime Prevention*. Routledge.

Gatt C (2023) Decolonizing scholarship? Plural onto/epistemologies and the right to science. *Frontiers in Sociology* 8: 1297747.

Glanzberg M (2023) Truth. In: Zalta EN and Nodelman U (eds) *The Stanford Encyclopedia of Philosophy*. Fall 2023. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/fall2023/entriesruth/ (accessed 15 May 2025).

House ER (2014) Origins of the Ideas in Evaluating with Validity. *New Directions for Evaluation* 2014(142): 9–15.

House ER and Howe KR (1999) *Values in Evaluation and Social Research*. 1st edition. Thousand Oaks, Calif: SAGE Publications, Inc.

Khosrowi D (2019) TRADE-OFFS BETWEEN EPISTEMIC AND MORAL VALUES IN EVIDENCE-BASED POLICY. *Economics & Philosophy* 35(1): 49–78.

Krauss A (2021) Assessing the Overall Validity of Randomised Controlled Trials. *International Studies in the Philosophy of Science* 34(3): 159–182.

Kuhn TS (1979) *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago, IL: University of Chicago Press. Available at: https://press.uchicago.edu/ucp/books/book/chicago/E/bo5970650.html (accessed 15 May 2025).

Kushner S and Stake R (2025) Breakthroughs, advocacies and a return to validity in programme evaluation. *Evaluation* 31(1). SAGE Publications Ltd: 7–21.

Lusk G and Elliott KC (2022) Non-epistemic values and scientific assessment: an adequacy-for-purpose view. *European Journal for Philosophy of Science* 12(2): 35.

McMullin E (1982) Values in Science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1982. [University of Chicago Press, Springer, Philosophy of Science Association]: 3–28.

Parker WS (2009) Confirmation and Adequacy-for-Purpose in Climate Modelling. *Proceedings of the Aristotelian Society, Supplementary Volumes* 83. [Oxford University Press, The Aristotelian Society]: 233–249.

Parker WS (2020) Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science* 87(3): 457–477.

Patton MQ (2010) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. Guilford Press.

Pawson R (2006) *Evidence-Based Policy: A Realist Perspective*. SAGE.

PIR Methods (n.d.). Available at: https://www.3ieimpact.org/pir-methods (accessed 4 June 2025).

Quadrant Conseil (2017) Impact Tree. Available at: https://www.quadrant-conseil.fr/ressources/impacttree.html (accessed 15 May 2025).

Regt HW de (2017) *Understanding Scientific Understanding*. Oxford Studies in Philosophy of Science. Oxford, New York: Oxford University Press.

Reichardt CS (2022) The Counterfactual Definition of a Program Effect. *American Journal of Evaluation* 43(2): 158–174.

Rorty R (1998) *Truth and Progress: Philosophical Papers*. 1st ed. Cambridge University Press. Available at: https://www.cambridge.org/core/product/identifier/9780511625404/type/book (accessed 15 May 2025).

Rudner R (1953) The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science* 20(1). [The University of Chicago Press, Philosophy of Science Association]: 1–6.

Saunders M (2000) Beginning an Evaluation with RUFDATA: Theorizing a Practical Approach to Evaluation Planning. *Evaluation* 6(1). SAGE Publications Ltd: 7–21.

Saunders M (2012) The use and usability of evaluation outputs: A social practice approach. *Evaluation* 18(4): 421–436.

Schwandt TA and Gates EF (2021) *Evaluating and Valuing in Social Research*. 1st edition. New York ; London: Guilford Press.

Scriven M (2007) The Logic of Evaluation. In: *OSSA Conference Archive*, 2007.

Steel D (2010) Epistemic Values and the Argument from Inductive Risk*. *Philosophy of Science* 77(1). [The University of Chicago Press, Philosophy of Science Association]: 14–34.

Stern E, Stame N, Mayne J, et al. (2012) *Broadening the range of designs and methods for impact evaluations*. April. Institute for Development Studies. Available at: http://repository.fteval.at/id/eprint/126 (accessed 15 May 2025).

Stirling A (2015) Developing 'Nexus Capabilities': towards transdisciplinary methodologies. In: *STEPS Centre*, Sussex, June 2015. Available at: https://steps-centre.org/publication/developing-nexus-capabilities-towards-transdisciplinary-methodologies/ (accessed 24 May 2025).

Wing C and Cook TD (2013) Strengthening the Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison. *Journal of Policy Analysis and Management* 32(4): 853–877.