African Evaluation Journal
ISSN: (Online) 2306-5133, (Print) 2310-4988

Evaluability perspectives: An empirical investigation of programme evaluability in different practice contexts

Authors:
Adiilah Boodhoo[1]
Joha Louw-Potgieter[1]

Affiliations:
[1]School of Management Studies, University of Cape Town, Cape Town, South Africa

Corresponding author:
Joha Louw-Potgieter,
joha.louw-potgieter@uct.ac.za

Dates:
Received: 30 Aug 2019
Accepted: 09 Dec. 2019
Published: 17 Feb. 2020

How to cite this article:
Boodhoo, A. & Louw-Potgieter, J., 2020, 'Evaluability perspectives: An empirical investigation of programme evaluability in different practice contexts', *African Evaluation Journal* 8(1), a434. https://doi.org/10.4102/aej.v8i1.434

Background: The empirical literature on programme evaluability is sparse and little is known about how evaluators operationalise prescriptive articulations of evaluability.

Objectives: In this study, we explore inductively and comparatively how evaluators practising in different contexts (i.e. high-income or middle-income countries, with or without mature evaluation cultures) operationalise programme evaluability.

Method: We administered the Q-sort method to a geographically dispersed expert sample and systematically identified evaluability perspectives that are unique to and shared across different evaluator cohorts. Valid responses from evaluators recruited from the United States of America (USA) ($n = 86$), the United Kingdom ($n = 26$), Brazil ($n = 79$) and South Africa ($n = 38$) were analysed using Q factor analysis.

Results: Four empirically distinct perspectives could be characterised meaningfully, two of which (labelled as *theory-driven* and *utilisation-focused*) were shared by most evaluators in our sample.

Conclusion: Implications for cross-border collaborations as well as viable strategies to reconcile divergent perspectives that could emerge within evaluability assessment teams are discussed.

Keywords: evaluability perspectives; evaluability assessments; evaluation context; cross-border evaluation; Q-sort method.

Introduction

Despite recent publications on evaluability assessments (EAs) (e.g. Davies & Payne 2015; Trevisan & Walser 2014; Walser & Trevisan 2016) and widespread deliberation around the topic on online evaluation blogs and forums, such as EVALTALK and EA365, the concept of evaluability remains elusive. To date, there is no empirically validated method for conducting sound EAs, limited systematic evidence on their effectiveness in general and no empirical characterisation of what evaluators think evaluability looks like.

Prescriptive articulations, in the form of EA process models or evaluability checklists, have been criticised for their lack of operational specificity (Davies & Payne 2015). The role of prescriptive articulations, in general, appears to be a contentious matter, with some distinguished evaluators, such as Michael Scriven, dismissing their importance, and others viewing them as fundamental to our professional entity (Donaldson & Lipsey 2006). Studies on evaluation practice tell an interesting story: evaluators do not necessarily conform to prescriptive articulations in their everyday practice but draw on their implicit and pragmatic theories, shaped by factors such as experience, training and practical reasoning (Christie 2003a; Shadish & Epstein 1987). Consequently, it seems important to investigate systematically what 'folk theories' exist around different areas of our practice (Christie 2003b:92). This article reports on an exploratory study of 'folk theories' that exist in a critical and under-investigated area of our practice, namely, programme evaluability.

Programme evaluability and the evaluability assessment method and process: Some caveats

Before launching into the methodological aspects of our investigation, we need to address a few caveats. Firstly, programme evaluability is a dated concept (notions of evaluability emerged in the early 1970s), with inconsistent operationalisations articulated in evaluability checklists and

guidance material (Davies 2013). These operationalisations are often unqualified (i.e. there are no explicit weightings tied to the different evaluability criteria articulated, conveying the mistaken assumption that they are all of equal importance). However, there appears to be some consensus amongst international development agencies on the general meaning of the term 'evaluability'. The following working definition from the Organisation for Economic Cooperation and Development's (OECD) Development Assistance Committee (DAC) is widely quoted and used: 'the extent to which an activity or project can be evaluated in a reliable and credible fashion' (OECD-DAC 2010:21). Embedded in this definition is the notion of feasibility and the sentiment that we can 'literally evaluate anything, at least in some way, at some level, and at a certain cost' (Finckenauer, Margaryan & Sullivan 2005:266), but not necessarily in a reliable and credible manner. This definition can be further expanded to capture the evolving purposes of EAs since the inception of the method in the late 1970s – from determining programme readiness for summative evaluations (Wholey 1979) to a much broader scope, which includes ensuring that relevant and technically feasible evaluations are conducted, maximising evaluation utility, building evaluation capacity and determining the feasibility of implementing the desired evaluation design (Davies 2013; Leviton et al. 2010). This is a difficult undertaking given that there is no consensus on what EAs should achieve. One could argue, however, that a comprehensive definition of evaluability should, at the very least, tap into some (if not all) of the utility, feasibility, propriety, accuracy and accountability standards of the Joint Committee on Standards for Educational Evaluation (JCSEE) (Yarborough et al. 2019), and have at its core a recognition of the desired type of evaluation.

Secondly, EA process models that prescribe the procedural steps that underlie EAs are often discussed in isolation of the operationalisations or evaluability criteria embedded in evaluability checklists. The development of EA process models appears to be within the remit of evaluation theorists, whilst international development agencies such as the United Nations Fund for Women (UNIFEM) concentrate on the development of evaluability checklists applicable to their programmes. For instance, the UNIFEM checklist consists of 17 questions framed around evaluability parameters of programme design, availability of information and conduciveness of the context. Existing EA models represent adaptations of Wholey's (1979) original eight-step model, which he later refined into six-step model (Wholey 2004), articulating the following iterative process: (1) involving intended users of the evaluation, (2) clarifying the intended programme, (3) exploring programme reality, (4) reaching agreement on any required programme changes, (5) exploring alternative evaluation designs and (6) agreeing on evaluation priorities and intended uses of information. Examples of adaptations include Kaufman-Levy and Poulin's (2003) five-task model for EAs, Smith's (1981) 10-step EA model and Thurston and Potvin's (2003) seven-step framework. Despite the proliferation of EA models, with reasonably well-defined procedural steps, there is no agreed-upon operationalisation of what a well-executed EA process should look like (Trevisan 2007; Watts & Washington 2016).

This brings us to our third caveat: evaluability is not an absolute condition. Rather, it occurs along a continuum from more to less evaluable. A categorical judgement of evaluability is not only restrictive, but also at odds with many evaluation approaches and the broader scope of EAs nowadays (Davies 2013).

Our aims and investigative approach

The empirical literature on programme evaluability is sparse and little is known about how evaluators operationalise prescriptive articulations of evaluability. The purpose of this article is to refocus our attention on the concept and its empirical operationalisation(s). Our stance is that we need sharper insight into the 'what' of EA, from evaluators' perspectives, before deliberating on issues relating to the EA method or process (i.e. the 'how' of EAs) – most of which stem from rationalistic assumptions that underlie the EA process (Dahler-Larsen 2012). As such, our study aims to explore inductively whether or not evaluators have a common perspective towards evaluability, despite the ambiguous articulation of the concept of evaluability in the literature. More specifically, it addresses the following research question:

> Do evaluators share a common perspective towards evaluability? If not, what perspectives can be empirically identified?

We draw on Kundin's (2010) framework for studying evaluators' practice decisions. The framework isolates three key elements that might shape an evaluator's decision-making process: situation awareness, practical reasoning and reflection in action.

What we are interested in capturing in our study is working logic – a component of practical reasoning, which goes hand in hand with general logic. The first type of logic 'specifies the game and the rules of the game that one is playing when conducting an evaluation in any field' (Fournier 1995:17).

The second type of logic, also referred to as logic in use or reconstructed logic, represents the different operationalisations of general logic. In other words, they represent the variations in application of general logic in practice.

In our view, EA is a systematic 'game' with abstract 'rules' and room for variations in application. These variations can be isolated empirically by reconstructing evaluators' working logic of both the EA method or process (i.e. the procedural steps that underlie an EA) and the evaluability criteria to be prioritised in an EA. We are interested in the latter.

We take the study one step further by contrasting the evaluability perspectives of four different evaluator cohorts practising in different contexts. Theorists have long recognised the role of context in shaping evaluation practice

(e.g. Stake 1990). Despite the growing emphasis on the role of context in evaluation, there is no unified understanding of what context means and how exactly it influences evaluation practice (Dahler-Larsen & Schwandt 2012). We are interested in the practice of evaluation in different countries. As such, a secondary aim of our study is to explore comparatively how evaluators from selected countries operationalise programme evaluability.

We suspect that evaluators who practise in middle-income countries with emerging evaluation cultures might hold different evaluability perspectives than those who practise in upper-income countries with mature evaluation cultures. It is imperative that we subject this assumption to empirical testing and deliberate on its implications as evaluation is becoming 'more global and more transnational' and 'problems and programs that we are called upon to evaluate today often extend beyond the boundaries of any one nation, any one continent, or even one hemisphere' (Chelimsky & Shadish 1997:xii). The practice contexts (countries) we investigate in our study are described in more detail under the heading 'Participants'.

Method

Design

A descriptive design was used in this study as we were primarily concerned with the collection and description of cross-sectional data relating to participants' perspectives of programme evaluability.

Measures

Our focus was to analyse evaluators' subjective operationalisations of programme evaluability. We did so by applying the Q-sort method (Stephenson 1935), a procedure that facilitates the systematic study of participant subjectivity and the different accounts that people construct (Cross 2005). Whilst Q studies have often been criticised as small-sample investigations of unknown reliability, such scepticisms have been discredited by authors such as Eghbalighazijahani, Hine and Kashyap (2013) and Brown (1993). We also drew on the literature to inform our decisions regarding the relative sizes of the P set and the Q set, and the condition of instruction, and as such are convinced that the Q-sort method, as applied in our study, is sufficiently robust for examining subjective operationalisations of programme evaluability.

In a Q study, participants are instructed to sort a set of randomly ordered statements relating to a specific topic into a subjectively meaningful pattern based on their individual preference or judgement – a procedure called Q sorting. The individual rankings are then subjected to a Q factor analysis. A defining feature of the Q-sort method is that statements relating to the same domain are not analysed individually but in the context of other equally relevant statements.

We chose to use the Q-sort method in this study as it is particularly well suited to studying the phenomena 'in which there are numerous ideals present in a reality where only a limited number of ends or means can be realistically pursued' (Thompson 1998:1). This is particularly evident in an evaluation context, where practical realities constrain evaluators to prioritise a set of evaluability criteria at the expense of others. The Q-sort method would allow us to simulate prioritisation patterns of participating evaluators and derive evaluability perspectives that embody variations in working logic.

We firstly examined the existing opinions and perspectives around the topic of evaluability by means of an exhaustive literature review (Boodhoo 2016). In Q methodology, this step refers to the definition of the concourse. In a second step – the development of the Q set – we extracted and synthesised the evaluability criteria most commonly cited in the concourse and categorised them under three dimensions: (1) programme characteristics or structural features, (2) methodological or logistical requirements and (3) stakeholder characteristics (Appendix 1). The seminal writings of Wholey (1979), Schmidt, Scanlon and Bell (1979) and Nay and Kay (1982), as well as the more recent works of Davies (2013) and Dahler-Larsen (2012), laid the foundation of this exercise because of their systematic and comprehensive coverage of 'evaluable models' and evaluability parameters (unlike other recent or previous EA-related publications, which tend to focus on the application of EA process models or evaluability criteria established in seminal writings).

We then standardised the formulation of the Q set by drawing on conceptual definitions specified by Rossi, Lipsey and Freeman (2004) and made further refinements based on common usage and pragmatism. This process culminated in a set of refined, standardised evaluability criteria, formulated in terms of 19 Q statements (see Table 1), which could be

TABLE 1: Standardised evaluability statements.

Category	Evaluability statements
Programme characteristics	
Programme goals and outcomes	Programme goals are clearly specified
	Programme outcomes are realistic
	Programme outcomes are measurable
	Stakeholders agree on programme goals
Programme data	Programme data are adequate
	Programme data are reliable
	Programme data are easily accessible
Programme theory	Programme theory is explicitly stated
	Programme theory is plausible
Programme design	Service delivery is clearly defined
	Target beneficiaries are clearly defined
Programme implementation	Programme is implemented as intended
Stakeholder characteristics	Stakeholders are willing to collaborate with the evaluator
	Stakeholders have authority to act on evaluation findings
	Stakeholders are transparent about the purpose of the evaluation
Logistical requirements	Budget is adequate for the evaluation
	Time frame is adequate to complete the evaluation
	Type of evaluation required (process, outcome or impact) is feasible
	Required evaluation methodology is feasible

classified into one of the five ranking categories (ranging from *not at all important* to *essential*), according to their assigned importance.

Participants

In the Q-sort method, participants (referred to as the P set) are purposively selected and expected to possess clear and varied viewpoints on the topic under investigation – in our case, evaluators were recruited from Brazil, South Africa (SA), the United Kingdom (UK) and the United States of America (USA). These four practice contexts were selected based on their current socio-economic standing and the maturity of their evaluation cultures. We also capitalised on the relative ease of accessing our target population through the membership database of four professional associations (the Brazilian Monitoring and Evaluation Network, the South African Monitoring and Evaluation Association, the UK Evaluation Society and the American Evaluation Association) and our professional networks in each country.

Each country of interest could be categorised as a high-income or middle-income country (using the 2018 World Bank country classification system), with emerging or mature evaluation cultures (following an in-depth qualitative analysis of the historical trajectory of programme evaluation in each country, including its emergence and development). We recognise the inherent difficulty in ranking countries based on the maturity of their evaluation culture – a broad concept, with no perfect operationalisation (Dahler-Larsen & Boodhoo 2019; Jacob, Speer & Furubo 2015). We simply contend that Brazil and SA and the UK and USA could be considered, in broad terms, as comparable practice contexts. For example, Brazil and SA (both classified as developing countries) have similar historical, social and economic trajectories. Programme evaluation also emerged around the same time in both countries (in the 1990s), as part of their re-democratisation process (Abrahams 2015; Henriques et al. 2010). The firm and uniform institutionalisation of evaluation is, however, still underway in these two countries (Goldman et al. 2018; OECD 2017).

The USA and the UK, on the other hand, belong to a cohort of developed countries considered as pioneers in the development of public policy evaluation, with an evaluation tradition that dates back to the early 1960s (Gray & Jenkins 2002; Rist & Paliokas 2002). The USA and UK have also been characterised as having a high degree of evaluation culture maturity (Jacob et al. 2015).

Participants from the USA (n = 86) were highly experienced evaluators, with 28.7% having between 11 and 15 years of evaluation experience, and 30.9% holding a PhD degree in evaluation. Participants from the UK (n = 26) were also experienced evaluators, with 36.7% having between 6 and 10 years of evaluation experience. Most of these participants (63.3%), however, had not received any formal training in evaluation. The Brazilian and South African cohorts (n = 79; n = 38) were the least experienced, with 40.7% and 33.3%, respectively, having between 1 and 5 years of evaluation experience. Most of these evaluators were either self-educated or had completed a short course certificate in evaluation.

Our small sample size was not considered problematic, given our study aims and method of choice. The Q method attempts to isolate subjective structures in the data and the extent to which these are similar or dissimilar, rather than calculating the percentage of the sample or population that adheres to them (Eghbalighazijahani et al. 2013). The issue of generalisability is therefore not relevant here. In fact, there is no clear-cut rule on the minimum number of participants to include in a P set (Dziopa & Ahern 2011).

In addition, the focus of the method is not on the constructors (i.e. the participants) but on the accounts that they construct. As would be the case with any other topic, we expect only a limited number of distinct perspectives to exist on evaluability, and we are confident that our carefully selected P set is adequate to reveal these perspectives.

Procedure

We collected our data over a period of 2 months following ethical clearance from the Ethics in Research Committee of the Faculty of Commerce, University of Cape Town, and the translation of our study materials into Portuguese for the Brazil cohort. We used an online data collection platform, given our geographically dispersed sample. Results of self-administered, electronic-based Q sorts have been found to be consistent with traditional methods of administration (Reber, Kaufman & Krop 2000).

We used a free-sort condition of instruction, whereby participants were instructed to distribute iteratively randomly ordered Q statements across five ranking categories (representing different gradients of importance), until all 19 Q statements had been sorted. We opted against a forced-choice condition of instruction (which 'forces' the placement of Q items in a fixed and specified distribution), in light of the reported difficulty experienced by participants in a small pilot study we ran prior to the four-country study. Whilst a free-sort condition of instruction will inevitably yield variations in the 'shapes' of item distributions, this is not expected to affect our overall factor analytic solution (Watts & Stenner 2012).

After completing the Q-sort, participants had to respond to 12 items relating to their current involvement in evaluation, employment setting, highest academic qualification, type of training in evaluation, level of experience in conducting different types of evaluations, practice context, as well as the number of evaluations that they conducted in the last 5 years.

Data analysis

We applied Q factor analysis (with principal component analysis as a method of factor extraction) across our four data sets to identify (1) dominant evaluability perspectives that may be unique to each cohort of evaluators and (2) the

specific criteria that characterise those perspectives. What distinguishes Q factor analysis from conventional factor analysis is not the mathematics of the factor analytic process but the organisation of the raw data matrix, with rows representing Q statements and columns representing respondents' Q sorts.

We did not identify any systematic sorting pattern that warranted the exclusion of specific Q sorts for the US, UK and South African evaluator cohorts. Three problematic cases were, however, identified amongst Brazilian respondents and thus deleted from subsequent analyses: one respondent allocated all 19 Q statements to the *not at all important* category, and two respondents used only the first two categories *not at all important* and *quite unimportant* to distribute the Q statements (distribution ratio of 18:1 and 17:2, respectively).

We first confirmed the factorability of the inter-correlation matrixes – all factored entities had a correlation of at least 0.3 with multiple other entities, and all communalities were well above 0.5. Only factors with eigenvalues greater than 1, and positioned to the left of the inflexion point in the scree plot, whilst cumulatively explaining at least 60% of the total variance, were retained for Varimax factor rotation.

We deleted the following cases after examining the initial rotated factor matrixes: (1) 32 respondents with no significant factor loadings on any factors (we applied a 0.50 cut-off point), (2) 11 respondents with multiple significant cross loadings and (3) 13 respondents who were not adequately accounted for by the factor solution (i.e. respondents with significant factor loadings but with communalities < 0.50). We re-ran the analysis and deleted problematic cases iteratively until a satisfactory factor solution with pure factor loadings (indicating well-defined perspectives) and high factor reliability was obtained for each evaluator cohort. The reliability of a factor is determined by the number of respondents that define it. According to Brown (1993), five respondents are sufficient to obtain a clear reading of the perspective embodied in a given factor. Any additional respondents would only marginally clarify the picture.

We then examined the factor scores, which indicate the importance of each Q statement in defining a given rotated factor. In line with Thompson (1998), we used a cut-off score of -1 to identify the most important and least important evaluability criteria associated with a particular factor.

As a final step, we interpreted the different factors or viewpoints that emerged from the Q analysis. In the absence of a set strategy for factor interpretation in the Q method literature, we applied the interpretative framework proposed by Watts and Stenner (2012) to arrive at a systematic and holistic interpretation of each factor. We developed a 'crib sheet' to organise the Q statements based on the size and rank order of their associated factor scores. The following distinctions were made: (1) statements with the highest rankings in each factor array, (2) statements with the lowest rankings in each factor array, (3) statements with higher rankings in a given factor array compared to other factor arrays and (4) statements with lower rankings in a given factor array compared to other factor arrays. We considered the positioning of each statement and applied the logic of abduction to generate iteratively the overall story underlying the various statement rankings and derive preliminary hypotheses that could account for a particular item configuration or factor array.

We also selected an appropriate label to represent each factor by examining the Q statements that distinguished them. Whilst factor labelling is not a methodological requirement, it conveys in a parsimonious manner what distinguishes factors from one another. Statements with positive factor scores (i.e. those characterised as essential) were given more weight in factor labelling. Whilst this approach does not in any way capture the complexity of a given viewpoint, we strove to label each perspective in a manner that best integrates all the distinguishing statements associated with it.

The final interpretation of each factor is presented in narrative form. The relevant statements were linked together to create a unified account of the viewpoint embodied in each of the factors identified.

Ethical consideration

Ethical clearance was obtained from the Ethics in Research Committee, Faculty of Commerce, University of Cape Town on 17 March 2014, no clearance number was issued at the time.

Results

Table 2 reflects that at least two dominant perspectives, with high factor reliability, could be meaningfully interpreted for each evaluator cohort (relevant crib sheets and factor scores can be requested from the first author).

Evaluability perspectives of evaluators recruited from the United States of America

Factor 1 reflects a perspective that favours an explicit change logic and implementation fidelity but minimises the importance of logistical imperatives and stakeholder collaboration, authority and transparency. It would seem that the underlying focus is on opening the black box of evaluation and making the underlying assumptions and implementation of the programme activities clear. This perspective could be construed as the essence of a theory-driven evaluation approach and was labelled as such. Twenty-two respondents in our final USP set (n = 56) were most associated with factor 1.

Factor 2 seems to reflect evaluators' concern with mechanisms that support the utilisation of results by intended users: stakeholder transparency, authority and consensus. We therefore labelled this perspective as *utilisation-focused*. Twenty-two respondents were most associated with factor 2.

TABLE 2: Summary of results.

Evaluator cohort	No. of factors with high factor reliability	% of variance explained by each factor	No. defining each factor	Factor label
United States of America (USA)	4	Factor 1: 24.5	22	Theory-driven
		Factor 2: 23.2	22	Utilisation-focused
		Factor 3: 8.9	-	Undefined
		Factor 4: 8.8	-	Undefined
United Kingdom (UK)	2	Factor 1: 31.1	8	Theory-driven
		Factor 2: 25.0	5	Theory-driven and utilisation-focused
Brazil	4	Factor 1: 22.3	14	Theory-driven
		Factor 2: 14.9	9	Utilisation-focused
		Factor 3: 13.0	-	Undefined
		Factor 4: 10.4	6	Implementation-focused
South Africa (SA)	2	Factor 1: 27.8	10	Theory-driven
		Factor 2: 16.6	6	Utilisation-focused

The item configuration of both factors 3 and 4 is not related to any explicit notion of evaluation practice, thus making the characterisation of these perspectives problematic. Both factors also explained a relatively low proportion of variance (8.9% and 8.8%, respectively) and were defined by a small number of respondents ($n = 6$).

Evaluability perspectives of evaluators recruited from the United Kingdom

The first factor extracted for the UK cohort reflects a perspective that minimises the importance of stakeholder collaboration, authority and transparency, and certain logistical imperatives. Here, the underlying focus is on the ability to measure implementation fidelity and explain why a programme worked or did not work. This perspective is similar to the first factor identified for the US cohort. As such, we decided to label it as *theory-driven*. Eight respondents from our final UKP set ($n = 13$) were most associated with this perspective.

The second factor reflects a perspective that emphasises the need for (1) a logic model that articulates realistic and measurable outcomes and (2) stakeholders' transparency and authority. An explicit programme delivery plan and certain logistical requirements are not of high priority. There seems to be a dual focus on programme theory, and the necessary conditions for utilisation of evaluation findings. This factor reflects a combined *theory-driven* and *utilisation-focused* perspective and was labelled as such. Five respondents were most associated with this perspective.

Evaluability perspectives of evaluators recruited from Brazil

Fourteen respondents (out of the 36 in the Brazil P set) were most associated with the first factor. This perspective emphasises the need for an explicit and plausible theory of change, which operationalises clearly specified and agreed-upon programme goals. Evidence that the programme has been implemented with fidelity is also of high priority. It would seem that the underlying focus is on mechanisms that support the change process or programme success and the ability to explain why the programme worked or did not work. This bottom-up approach mirrors a theory-driven approach to evaluation. We decided to label this perspective as *theory-driven*.

Nine respondents were most associated with the second factor. Evaluators who shared this perspective prioritised stakeholder transparency, authority and consensus and fidelity of implementation. Issues pertaining to data collection, evaluation design and programme theory were assigned less importance. Given that the overall focus is on mechanisms that support the utilisation of findings, we labelled this perspective as *utilisation-focused*.

For factor 3, we could not meaningfully reconcile the entire item configuration in terms of evaluation practice and therefore decided not to retain it.

For factor 4, the emphasis was on the specification and proper implementation of the service delivery plan, availability and accessibility of data and sufficient budget to conduct the evaluation. Data quality, plausibility of the change logic and consensus on programme goals are of lower priority for the six evaluators who shared this perspective. The underlying focus appears to be on the minimum requirements to measure implementation fidelity and it was labelled as *implementation-focused*.

Evaluability perspectives of evaluators recruited from South Africa

Ten respondents (out of 16 in the South African P set) were most associated with the first factor. Evaluators who shared this perspective seemed to prioritise (1) an explicit and plausible theory, articulating realistic and measurable outcomes for a specific target population; and (2) data for measuring implementation fidelity. Taken together, the underlying focus appears to be on opening the black box of evaluation, and the ability to explain why the programme worked or did not work. This perspective resonates with a theory-driven approach to evaluation and was labelled as such.

Six respondents were most associated with the second factor. Evaluators who shared this perspective prioritised stakeholder transparency, authority, consensus and collaboration. Issues pertaining to evaluation design, data

collection and evaluation time frame are of less priority. Given that the overall focus is on mechanisms that support the utilisation of findings, we labelled this perspective as *utilisation-focused*.

Discussion

Our main research question focused on whether evaluators share a common perspective towards evaluability, and if not, what perspectives can be identified empirically? Four empirically distinct perspectives emerged from the data, suggesting that evaluators may approach EAs differently. The first perspective (theory-driven) was shared by all four evaluator cohorts and can thus be considered as the most dominant perspective (52 evaluators defined this particular perspective). The second perspective (utilisation-focused) was shared by at least one group of evaluators from the USA (n = 22), Brazil (n = 9) and SA (n = 6). The *implementation-focused* perspective and the combined *theory-driven* and *utilisation-focused* perspective were unique to the Brazilian and UK cohorts, respectively.

Several general conclusions can be drawn. Firstly, the perspectives of evaluators within each evaluator cohort were quite different, with evaluators from Brazil having the most divergent perspectives on evaluability. Secondly, the finding that certain evaluability perspectives were shared by all evaluator cohorts suggests that the views of a select group of evaluators were compatible, even if they did not practise in the same context.

Divergent or multiple evaluability perspectives within evaluator cohorts: Reasons, implications and solutions

Here, we deliberate on four questions that stem logically from the first finding of this study and present our preliminary thoughts.

Why were divergent or multiple evaluability perspectives identified within each evaluator cohort?

One might argue that the existence of competing perspectives towards evaluability indicates that the evaluation community does not have a clear and collective understanding of this particular construct. This is a reasonable argument given that (1) there is a lack of consensus on the working definition of evaluability in the literature, and (2) the extant grey literature is fraught with debates over the fundamentals of evaluability (e.g. some evaluators question the need to assess evaluability, given that any programme can be subjected to some form of evaluation, whilst others question whether or not evaluability can be measured).

Should we work towards a unified perspective on evaluability?

We contend that multiple or divergent perspectives could be adaptive as long as there is some agreement on the fundamentals (Cooper 2014). Very few professions are characterised by practitioners who rigidly adhere to one distinct perspective on an issue, and according to Shadish and Epstein (1987), there is no obvious motivation for programme evaluators to be any different. Whilst we need to have a unified perspective on evaluability (given the range of acceptable evaluation approaches, purposes and methods), we need to deliberate on the intricacies of each perspective. Whilst we appreciate the difficulty of this task, it would be useful to accumulate empirical knowledge regarding the practice of evaluators with different evaluability perspectives (e.g. one can investigate the inherent challenges associated with each perspective and how evaluators overcome these challenges when conducting EAs), with a view to develop descriptive theories of evaluation practice.

What are the implications of having multiple or divergent evaluability perspectives on our discipline and practice?

This question can be addressed by examining the pragmatic challenges of having an evaluation community characterised by multiple or divergent perspectives towards programme evaluability. One can reasonably argue that the practice of evaluators who share a common perspective is informed by a common set of underlying values. They are most likely to agree on a number of fundamental issues, such as what counts as good practice, what questions are worth investigating and what methods are to be used. Divergent perspectives towards evaluability could therefore have a number of negative implications on collaborative work. For example, more time might be needed to resolve fundamental differences in approaches or concept definition, interdependent activity might become more difficult to coordinate and efficiency might be compromised by greater task uncertainty. Many evaluations are collaborative ventures, conducted by multidisciplinary teams. Whilst this is an untested notion, it is conceivable that evaluators with divergent perspectives towards evaluability might find it difficult to work collaboratively on an EA. For instance, evaluators with a theory-driven perspective might consider the assessment of stakeholders' level of instrumental authority and potential collaboration as a waste of valuable resources, whilst their counterparts who share a utilisation-focused perspective might lobby for such an assessment. The situation becomes even more challenging if an evaluation team consists of evaluators with fragmented perspectives on evaluability. The finding that 62% of participating evaluators in this study had fragmented perspectives lends credence to this possibility.

An evaluation community characterised by a lack of consensus on what constitutes an evaluable programme can stagnate in terms of skills and knowledge development, especially if there are minimal attempts to integrate or resolve fundamental differences across evaluators. At present, there is limited dialogue amongst evaluators with divergent perspectives on evaluability, and a thin empirical base to assess the merits of each perspective.

How can multiple or divergent evaluability perspectives be reconciled?

Here, we examine the strategies that contenders of other multi-paradigm disciplines (e.g. psychology) use to communicate and coordinate their actions. The first

applicable strategy is to fix the meaning of the term 'evaluability' as the concept is articulated in ambiguous and inconsistent terms in the literature (Trevisan 2007). The meaning of the term can be fixed by creating and validating prototypical examples of *unevaluable* programmes or programmes that are evaluable with difficulty. According to Cooper (2014):

> [*W*]hen the meaning of a term is fixed by pointing at an example of a kind, the fact that different [practitioners] may have different beliefs about things of that kind is irrelevant. Regardless of their different beliefs, all speakers talk about the same thing. (p. 99)

The second applicable strategy is to interact with evaluation stakeholders as a team throughout the EA process. This approach will ensure that all evaluators have direct access to the same contextual information, form a common frame of reference and make a joint decision about evaluability (task delegation or role differentiation might be counterproductive in this context; see Levesque, Wilson & Wholey 2001).

The third applicable strategy is to use the evaluability criteria imposed by external regulating bodies to encourage evaluators to set aside their conflicting perspectives for the purpose of collaborative work. The issue here is that different funding or development agencies have different evaluability checklists and scoring protocols. How do we decide which one to use? Even if we select the most comprehensive checklist, which evaluability dimension or criterion should be assigned the highest weighting, given that most checklists do not have pre-defined weights (Davies 2013), is a concern. In line with Davies and Payne (2015), we recommend an involved deliberation to ensure that the assigned weights are sensitive to the evaluation context, the specific type of evaluation to be conducted and the evaluation approach to be used.

The need for such an involved deliberation should be emphasised in evaluator training programmes. We believe that the conversation has to move beyond which evaluability perspectives to transmit as part of pre-service training programme or what is the best systematic approach for conducting an EA. Instead, we need to deliberate on how to train evaluators to resolve and integrate different perspectives that might emerge in collaborative undertakings – a plausible scenario given the results of this study. How do we transmit this type of practical wisdom which is typically acquired through extensive experience? In line with Trevisan (2004) and House (2015), we recommend any approach that would provide vicarious exposure to the intricacies of conducting EAs in real-world settings, and as part of an EA team. These include simulations with case descriptions, role-plays and practicum experiences. Prototypical examples of unevaluable programmes or programmes evaluable with difficulty could be used to elicit the evaluability perspectives espoused by EA teams in training.

Shared evaluability perspectives across evaluator cohorts: Characterisation and origins

Our second finding focused on the evaluability perspectives that were shared by all or most evaluator cohorts. We frame our discussion around the following three questions.

What principles underlie the main evaluability perspectives identified in this study?

The dominant perspective that emerged across all cohorts of interest was labelled as *theory-driven*. The hallmark of this perspective is its emphasis on unpacking 'programmatic black boxes and [explaining] how and why programs work (or fail to work) in different contexts and for different program stakeholders' (Astbury & Leeuw 2010:364). This perspective aligns with Epstein and Klerman's (2012) proposed logic model approach to evaluability. This approach requires the explicit specification of a programme's theory of change in the form of a 'falsifiable logic model' (Epstein & Klerman 2012:380). This model is an extension of the conventional logic model in that it contains extensive process-related detail, and quantitative benchmarks for programme operations and intermediate outcomes.

Epstein and Klerman's (2012) notion of an augmented logic model is at the core of recent conceptualisations of theory-driven evaluations and Pawson and Tilley's (1997) realist approach to evaluation. Theory-driven evaluations are driven by 'contextualized, comprehensive, [and] ecological program theory models' (Coryn et al. 2011:202) in an attempt to address the 'black box problem' (Astbury & Leeuw 2010:364). In realistic evaluations, it is not sufficient to link programmes causally to outcomes; identifying the underlying mechanisms that are triggered in particular contexts to produce the desired outcomes, as well as the structures that enable the intended mechanism of change, is key. Realistic evaluations attempt to unpack the context-mechanism-outcome configuration (CMOC) by developing and testing CMOC theories.

The second evaluability perspective that emerged consistently across three evaluator cohorts was labelled as *utilisation-focused* as it consisted of empirically supported factors that promote evaluation use. In their review of 41 empirical studies conducted over a 25-year period, Johnson et al. (2009) found that stakeholder involvement in the evaluation process and stakeholder–evaluator interaction and communication are key to maximise the evaluation use. This finding aligns with one of the main premises underlying Patton's (2008) utilisation-focused approach: evaluation is more likely to be used if primary intended users are involved in a meaningful manner in the evaluation process, feel ownership of the process and have a stake in the findings. Whilst stakeholder dynamics are not the only factors that have been linked to evaluation use, 23 out of 41 studies included in Johnson et al.'s (2009) review investigated this particular factor, and the bulk of these studies supported its relationship with other use factors.

The centrality of use in our evaluation practice is well recognised and has led to the development of participatory, stakeholder-based and collaborative approaches to evaluation (O'Sullivan 2012). The *utilisation-focused* evaluability perspective that emerged in this study aligns with these approaches. The evaluation literature is inundated by variants of both of these approaches (e.g. practical and transformative strands of participatory evaluation), all built on the underlying principle of extensive stakeholder involvement (Brandon & Fukunaga 2014).

What factors could have accounted for the emergence of these perspectives?

It is conceivable that these perspectives emerged because notions of 'unpacking the black box' (Astbury & Leeuw 2010:364), use and stakeholder involvement are firmly entrenched in our discipline and practice (Alkin & Taut 2003; Brandon & Fukunaga 2014). The origins of theory-driven evaluations can be traced back to the 1930s, more specifically to Tyler's early conceptualisation of the approach; however, it gained more prominence in the evaluation community with the publication of Chen's seminal book, *Theory-Driven Evaluations*, in 1990. Since then, this approach gained extensive coverage in the literature, and increased popularity amongst practitioners under the guise of theory-oriented evaluations, programme theory evaluation, intervening mechanism evaluation, programme theory-driven evaluation science and the like (Coryn et al. 2011).

Similarly, concern for evaluation use can be traced back to the 1960s (Alkin & Taut 2003), the early days of our profession. As evaluators, we have a long-standing interest in the intended and unintended influence of our work, as manifested by the conceptual, symbolic or instrumental use of evaluation findings, and the learning that occurs during the evaluation process (Johnson et al. 2009). This interest is central to our professional identity so much so that the concept of use or utilisation has been the subject of extensive deliberation in theoretical writings and is arguably the most well-researched area in the field. Evaluators continuously strive for a greater understanding of how evaluation use can be facilitated, and widely agree that stakeholder involvement plays a central role in this process. Stakeholder involvement is at the heart of our practice (Brandon & Fukunaga 2013) and underlies a number of formally endorsed principles for evaluators in the Global North, as well as firmly established and newly introduced approaches to evaluation, such as Hansen and Vedung's (2010) theory-based stakeholder evaluation approach.

It is also possible that evaluators who had well-defined perspectives on evaluability were predominantly theory-driven and utilisation-focused evaluators, or at the very least strong proponents of these approaches. This claim is in no way conclusive as the theoretical orientations of participating evaluators were not explored in this study – a gap that could be addressed by future research.

What are the implications of having shared evaluability perspectives across evaluator cohorts?

The emergence of compatible perspectives across evaluator cohorts might, for example, facilitate cross-border collaboration and dialogue amongst the US and Brazilian evaluators. This type of collaboration is particularly relevant in the context of cross-cultural evaluations and development evaluation (Chouinard & Cousins 2009), where 'a closer exchange of experiences between U.S. evaluation practitioners and their colleagues from developing countries could be mutually beneficial' (Bamberger 2000:101). Here, an EA team consisting of evaluators from geographically dispersed countries would be more cohesive if they shared a common understanding of what makes a programme evaluable.

Limitations

At least three methodological limitations of this study should be highlighted. Firstly, it is difficult to explain conclusively why theory-driven and utilisation-focused evaluability perspectives emerged consistently across evaluator cohorts, given the exploratory nature of our study. We merely provide plausible explanations that could account for the patterns that have emerged in our data and informed conjectures about their associated implications. Secondly, this study, like any study that relies predominantly on purposive and snowball sampling strategies, carries the risk of selection bias. There is a possibility that participating evaluators were inherently different from those who declined participation or withdrew from the study. Thirdly, we realise in retrospect that the exclusion criteria should have been more conservative to address the conceptual distinction between *country of residence* and *country of practice*, with the latter being more salient in the context of this study. Whilst the inclusion of the item 'where do you mostly do evaluation work?' served to distinguish evaluators who practised in developing countries from those who practised in developed countries, instances where an evaluator resided in a developing country but mostly practised in a developed country, or vice versa, were not accounted for.

Conclusions and directions for future research

Although predominantly descriptive, this exploratory study provides valuable insight into how four different cohorts of evaluators operationalise prescriptive theories of programme evaluability. As Smith (1993:240) remarked, 'if evaluation theories cannot be uniquely operationalized, then empirical tests of their utility become increasingly difficult'. We now have preliminary evidence that evaluators recruited from four different countries do not share a unified perspective towards evaluability. This result forces us to deliberate on how to train evaluators to resolve and integrate different perspectives that might emerge in collaborative undertakings. The results of our study also reinforced our initial stance that the 'what' of EA, from evaluators' perspectives, needs further development and articulation before deliberating on issues

relating to the 'how' of EAs (i.e. the method or process). There is scope to investigate the construct of evaluability more directly in subsequent studies. Future research could use scenario-based simulations to examine whether or not (1) evaluators reshape their operationalisations of evaluability depending on the presenting features of the evaluation 'context' and (2) evaluability decisions are tied to decisions to proceed with an evaluation. Such efforts would help us further clarify the meaning and use of this construct both in isolation and in relation to other potentially related constructs.

Acknowledgements

The authors would like to acknowledge and thank the following: the American Evaluation Association, the United Kingdom Evaluation Society, the South African Monitoring and Evaluation Association and the Brazilian Monitoring and Evaluation Network for providing access to their members; Veronica Pais and Marita Lindeque for translating the study instruments into Portuguese; and Katya Mauff and Alexa Heekes for assisting with interpreting the statistical analyses.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Authors' contributions

This manuscript was adapted from A.B.'s doctoral thesis for the degree of Doctor of Philosophy at the University of Cape Town. J.L.-P. served as the supervisor of the study. Both authors prepared a first draft of the manuscript for publication and worked on the final draft.

Funding information

The authors would like to acknowledge and thank the following funders: Adiilah Boodhoo – the University of Cape Town's Research Committee (URC) and the Postgraduate Centre and Funding Office; Joha Louw-Potgieter – the National Research Foundation's Incentive Fund for Rated Researchers (IFR150126113160).

Data availability statement

Data are available upon request from the first author.

Disclaimer

The views and opinions expressed in this article are the authors' own and not an official position of the institution or funder.

References

Abrahams, M.A., 2015, 'A review of the growth of monitoring and evaluation in South Africa: Monitoring and evaluation as a profession, an industry and a governance tool', *African Evaluation Journal* 3(1), 1–8. https://doi.org/10.4102/aej.v3i1.142

Alkin, M.C. & Taut, S.M., 2003, 'Unbundling evaluation use', *Studies in Educational Evaluation* 29(1), 1–12. https://doi.org/10.1016/S0191-491X(03)90001-0

Astbury, B. & Leeuw, F.L., 2010, 'Unpacking black boxes: Mechanisms and theory building in evaluation', *American Journal of Evaluation* 31(3), 363–381. https://doi.org/10.1177/1098214010371972

Boodhoo, A., 2016, *Evaluator characteristics and programme evaluability decisions: An exploratory study of evaluation practice in South Africa, Brazil, the United Kingdom, and the United States of America*, PhD Thesis, University of Cape Town, Cape Town.

Bamberger, M., 2000, 'The evaluation of international development programs: A view from the front', *American Journal of Evaluation* 21(1), 95–102. https://doi.org/10.1177/109821400002100108

Brandon, P.R. & Fukunaga, L.L., 2013, 'The state of the empirical research literature on stakeholder involvement in program evaluation', *American Journal of Evaluation* 35(1), 26–44. https://doi.org/10.1177/1098214013503699

Brown, S.R., 1993, 'A primer on Q methodology', *Operant Subjectivity* 16(3), 91–138.

Chelimsky, E. & Shadish, W.R., 1997, *Evaluation for the 21st century: A handbook*, Sage Publications Inc., Thousand Oaks, CA.

Chen, H.T., 1990, *Theory-driven evaluation*, Sage Publications Inc., Newbury Park, CA.

Chouinard, J.A. & Cousins, J.B., 2009, 'A review and synthesis of current research on cross-cultural evaluation', *American Journal of Evaluation* 30(4), 457–494. https://doi.org/10.1177/1098214009349865

Christie, C.A., 2003a, 'What guides evaluation? A study of how evaluation practice maps onto evaluation theory', *New Directions for Evaluation* 97, 7–36. https://doi.org/10.1002/ev.72

Christie, C.A., 2003b, 'Understanding evaluation theory and its role in guiding practice: Formal, folk, and otherwise', *New Directions for Evaluation* 97(Spring), 91–93. https://doi.org/10.1002/ev.79

Cooper, R., 2014, *Psychiatry and philosophy of science*, Routledge, New York.

Coryn, C.L., Noakes, L.A., Westine, C.D. & Schröter, D.C., 2011, 'A systematic review of theory-driven evaluation practice from 1990 to 2009', *American Journal of Evaluation* 32(2), 199–226. https://doi.org/10.1177/1098214010389321

Cross, R.M., 2005, 'Exploring attitudes: The case for Q methodology', *Health Education Research* 20(2), 206–213. https://doi.org/10.1093/her/cyg121

Dahler-Larsen, P., 2012, 'Evaluation as a situational or a universal good? Why evaluability assessment for evaluation systems is a good idea, what it might look like in practice, and why it is not fashionable', *Scandinavian Journal of Public Administration* 16(3), 29–46.

Dahler-Larsen, P. & Boodhoo, A., 2019, 'Evaluation culture and good governance: Is there a link?', *Evaluation* 25(3), 277–293. https://doi.org/10.1177/1356389018819110

Dahler-Larsen, P. & Schwandt, T.A., 2012, 'Political culture as context for evaluation', *New Directions for Evaluation* 135, 75–87. https://doi.org/10.1002/ev.20028

Davies, R., 2013, 'Planning evaluability assessment: A synthesis of the literature with recommendations', Report for the Department for International Development, Working Paper, no. 40, London.

Davies, R. & Payne, L., 2015, 'Evaluability assessments: Reflections on a review of the literature', *Evaluation* 21(2), 216–231. https://doi.org/10.1177/1356389015577465

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

Dziopa, F. & Ahern, K., 2011, 'A systematic literature review of the applications of Q-technique and its methodology', *European Journal of Research Methods for the Behavioral and Social Sciences* 7(2), 39–55. https://doi.org/10.1027/1614-2241/a000021

Eghbalighazijahani, M.A., Hine, J. & Kashyap, A., 2013, 'How to do a better Q-methodological research: A neural network method for more targeted decision making about the factors influencing Q-study', paper presented at the Irish Transport Research Network, Dublin, Ireland, 5–6 September.

Epstein, D. & Klerman, J.A., 2012, 'When is a program ready for rigorous impact evaluation? The role of a falsifiable logic model', *Evaluation Review* 36(5), 375–401. https://doi.org/10.1177/0193841X12474275

Finckenauer, J.O., Margaryan, S. & Sullivan, M.L., 2005, 'Evaluability assessment in juvenile justice a case example', *Youth Violence and Juvenile Justice* 3(3), 265–275. https://doi.org/10.1177/1541204005276267

Fournier, D.M., 1995, 'Establishing evaluative conclusions: A distinction between general and working logic', *New Directions for Evaluation* 68, 15–32. https://doi.org/10.1002/ev.1017

Goldman, I., Byamugisha, A., Gounou, A., Smith, L.R., Ntakumba, S., Lubanga, T. et al., 2018, 'The emergence of government evaluation systems in Africa: The case of Benin, Uganda and South Africa', *African Evaluation Journal* 6(1), 1–11. https://doi.org/10.4102/aej.v6i1.253

Gray, A. & Jenkins, B., 2002, 'Policy and program evaluation in the United Kingdom: A reflective state', in J.E. Furubo, R.C. Rist & R. Sandahl (eds.), *International atlas of evaluation*, pp. 129–153, Transaction Publishers, London.

Hansen, M.B. & Vedung, E., 2010, 'Theory-based stakeholder evaluation', *American Journal of Evaluation* 31(3), 295–313. https://doi.org/10.1177/1098214010366174

Henriques, A., Pinho, J., Azevedo, J.P. & Newman, J.L., 2010, 'The Brazilian monitoring and evaluation network: A report on the creation and development process', in G. Acevedo, K. Rivera, L. Lima & H. Hwang (eds.), *Challenges in monitoring and evaluation: An opportunity to institutionalize M&E systems*, pp. 164–171, The International Bank for Reconstruction and Development/The World Bank, Washington, DC.

Horst, P., Nay, J.N., Scanlon, J.W. & Wholey, J.S., 1974, 'Program management and the federal evaluator', *Public Administration Review* 34(4), 300–308.

House, E.R., 2015, *Evaluating: Values, biases and practical wisdom*, Information Age Publishing, Charlotte, NC.

Jacob, S., Speer, S. & Furubo, J.E., 2015, 'The institutionalization of evaluation matters: Updating the international atlas of evaluation 10 years later', *Evaluation* 21(1), 6–31. https://doi.org/10.1177/1356389014564248

Johnson, K., Greenseid, L.O., Toal, S.A., King, J.A., Lawrenz, F. & Volkov, B., 2009, 'Research on evaluation use a review of the empirical literature from 1986 to 2005', *American Journal of Evaluation* 30(3), 377–410. https://doi.org/10.1177/1098214009341660

Kaufman-Levy, D. & Poulin, M., 2003, *Evaluability assessment: Examining the readiness of a program for evaluation*, Juvenile Justice Evaluation Center, Justice Research and Statistics Association, Washington, DC.

Kundin, D.M., 2010, 'A conceptual framework for how evaluators make everyday practice decisions', *American Journal of Evaluation* 31(3), 347–362. https://doi.org/10.1177/1098214010366048

Levesque, L.L., Wilson, J.M. & Wholey, D.R., 2001, 'Cognitive divergence and shared mental models in software development project teams', *Journal of Organizational Behavior* 22(2), 135–144.

Leviton, L.C., Khan, L.K., Rog, D., Dawkins, N. & Cotton, D., 2010, 'Evaluability assessment to improve public health policies, programs, and practices', *Annual Review of Public Health* 31, 213–233. https://doi.org/10.1146/annurev.publhealth.012809.103625

Nay, J.N. & Kay, P., 1982, *Government oversight and evaluability assessment: It's always more expensive when the carpenter types*, Lexington Books, New York.

O'Sullivan, R.G., 2012, 'Collaborative evaluation within a framework of stakeholder-oriented evaluation approaches', *Evaluation and Program Planning* 35(4), 518–522. https://doi.org/10.1016/j.evalprogplan.2011.12.005

Organization for Economic Cooperation and Development (OECD), 2017, *Brazil's Federal Court of Accounts: Insight and foresight for better governance*, OECD Public Governance Reviews, OECD Publishing, Paris.

Organization for Economic Cooperation and Development's Development Assistance Committee (OECD-DAC), 2010, *Glossary of key terms in evaluation and results-based management*, viewed 03 January 2020, from https://www.oecd.org/dac/evaluation/glossaryofkeytermsinevaluationandresultsbasedmanagement.htm.

Patton, M.Q., 2008, *Utilization-focused evaluation*, Sage, Thousand Oaks, CA.

Pawson, R. & Tilley, N., 1997, *Realistic evaluation*, Sage, Thousand Oaks, CA.

Reber, B.H., Kaufman, S.E. & Cropp, F., 2000, 'Assessing Q-assessor: A validation study of computer-based Q sorts versus paper sorts', *Operant Subjectivity* 23(4), 92–209.

Rist, R.C. & Paliokas, K.L., 2002, 'The rise and fall (and rise again?) of the evaluation function in the US government', in J.E. Furubo, P.C. Rist & R. Sandahl (eds.), *International atlas of evaluation*, pp. 225–248, Transaction Publishers, London.

Rossi, P.H., Lipsey, M.W. & Freeman, H.E., 2004, *Evaluation: A systematic approach*, 7th edn., Sage, Thousand Oaks, CA.

Rutman, L., 1980, *Planning useful evaluations: Evaluability assessment*. Sage Publications, Beverly Hills, CA.

Schmidt, R.E., Scanlon, J.W. & Bell, J.B., 1979, *Evaluability assessment: Making public programs work better*, Monograph No. 14, Project SHARE, Rockville, MD.

Shadish, W.R. & Epstein, R., 1987, 'Patterns of program evaluation practice among members of the Evaluation Research Society and Evaluation Network', *Evaluation Review* 11(5), 555–590. https://doi.org/10.1177/0193841X8701100501

Smith, N.L., 1981, 'Evaluability assessment: A retrospective illustration and review', *Educational Evaluation and Policy Analysis* 3(1), 77–82. https://doi.org/10.2307/1163645

Smith, N.L., 1993, 'Improving evaluation theory through the empirical study of evaluation practice', *American Journal of Evaluation* 14(3), 237–242. https://doi.org/10.1177/109821409301400302

Stake, R.E., 1990, 'Situational context as influence on evaluation design and use', *Studies in Educational Evaluation* 16(2), 231–246. https://doi.org/10.1016/S0191-491X(05)80027-6

Stephenson, W., 1935, 'Correlating persons instead of tests', *Journal of Personality* 4(1), 17–24. https://doi.org/10.1111/j.1467-6494.1935.tb02022.x

Thompson, B., 1998, 'Using Q-technique factor analysis in *education program evaluations or research: An introductory primer', in Annual conference on research innovations in early intervention*, Charleston, SC, May 02, 1998, pp. 2–39.

Thurston, W.E. & Potvin, L., 2003, 'Evaluability assessment: A tool for incorporating evaluation in social change programmes', *Evaluation* 9(4), 453–469. https://doi.org/10.1177/135638900300900406

Trevisan, M.S., 2004, 'Practical training in evaluation: A review of the literature', *American Journal of Evaluation* 25(2), 255–272. https://doi.org/10.1177/109821400402500212

Trevisan, M.S., 2007, 'Evaluability assessment from 1986 to 2006', *American Journal of Evaluation* 28(3), 290–303. https://doi.org/10.1177/1098214007304589

Trevisan, M.S. & Walser, T.M., 2014, *Evaluability assessment: Improving evaluation quality and use*, Sage, Thousand Oaks, CA.

Tyler, R.W., 'Measuring the ability to infer', *Educational Research Bulletin* 9(17), 475–480.

Walser, T.M. & Trevisan, M.S., 2016, 'Evaluability assessment thesis and dissertation studies in graduate professional degree programs: Review and recommendations', *American Journal of Evaluation* 37(1), 118–138. https://doi.org/10.1177/1098214015583693

Watts, B.R. & Washington, H.M., 2016, 'Adaptation and use of a five-task model for evaluability assessment', *Journal of Multidisciplinary Evaluation* 12(27), 67–78.

Watts, S. & Stenner, P., 2012, *Doing Q methodological research: Theory, method and interpretation*, Sage, Thousand Oaks, CA.

Wholey, J.S., 1979, *Evaluation: Promise and performance*, Urban Institute, Washington, DC.

Wholey, J.S., 2004, 'Evaluability assessment', in J.S. Wholey, H.P. Hatry & K.E. Newcomer (eds.), *Handbook of practical program evaluation*, pp. 33–36, Josey-Bass, San Francisco, CA.

Wholey, J.S., 2010, 'Exploratory evaluation', in J.S. Wholey, H.P. Hatry & K.E. Newcomer (eds.), *Handbook of practical program evaluation* pp. 81–99, Josey-Bass, San Francisco, CA.

Yarborough, D.B., Shula, L.M., Hopson, R.K. & Caruthers, F.A., 2019, *The program evaluation standards: A guide for evaluators and evaluation users*, Sage, Thousand Oaks, CA.

Appendix starts on the next page →

Appendix 1

APPENDIX 1: Synthesis of evaluability parameters extracted from the evaluability concourse.

Evaluability parameter	Specification	Authors
Programme characteristics or structural features		
Programme objectives, goals, outcomes, expectations or effects	Well-defined or clearly specified	Wholey (1979, 2010), Horst et al. (1974), Rutman (1980), Davies (2013), Dahler-Larsen (2012)
	Realistic or plausible	Nay and Kay (1982), Schmidt et al. (1979), Stenberg and Wholey (1983)
	Measurable	Wholey (1979, 2010)
	Agreed upon	Wholey (1979, 2010), Nay and Kay (1982)
Programme data	Adequate	Davies (2013)
	Easily obtainable or accessible	Wholey (1979, 2010); Schmidt et al. (1979), Davies (2013)
	Reliable or valid	Schmidt et al. (1979), Davies (2013), Dahler-Larsen (2012)
Programme theory	Explicitly or consistently documented	Wholey (1979, 2010), Horst et al. (1974), Davies (2013)
	Plausible	Wholey (1979, 2010), Nay and Kay (1982), Rutman (1980), Davies (2013)
Programme design	Clearly defined intervention	Nay and Kay (1982), Schmidt et al. (1979), Rutman (1980), Dahler-Larsen (2012)
	Clearly defined target beneficiaries	Davies (2013)
Programme implementation	Implemented as intended	Wholey (1979, 2010), Rutman (1980), Stenberg and Wholey (1983)
Stakeholder characteristics		
Willingness	Willingness or availability to facilitate evaluation process	Horst et al. (1974), Rutman (1980), Davies (2013), Dahler-Larsen (2012)
Authority	Authority to facilitate evaluation process and act on evaluation findings	Nay and Kay (1982), Horst et al. (1974), Dahler-Larsen (2012)
Transparency	Clearly identified or agreed-upon information needs	Wholey (1979, 2010), Davies (2013), Stenberg and Wholey (1983), Dahler-Larsen (2012)
Methodological or logistical requirements		
Methodology	Feasibility of implementing desired methodology	Nay and Kay (1982), Rutman (1980), Davies (2013)
Evaluation type	Level of evaluation feasible	Stenberg and Wholey (1983), Davies (2013)
Budget	Adequate budget	Rutman (1980), Davies (2013), Stenberg and Wholey (1983)
Time	Adequate timeline	Rutman (1980), Davies (2013), Stenberg and Wholey (1983)