2

LEARNING OR COMPLYING? FINDING CONSENSUS ON THE PURPOSE OF EVALUATION

Caitlin Blaser Mapitsa & Khotso Tsotsotso

Introduction

Monitoring and evaluation systems are growing rapidly across the public sector in Africa, sparking enthusiastic debate about what makes these systems most appropriate and effective in design (Goldman, Byamugisha, Gounoug, Smith, Ntakumba, Lubanga, Sossou and Rot-Munstermann, 2018; Porter & Goldman, 2013). One of the factors contributing to this challenge is the disproportionate focus governments in the region place on compliance and monitoring (Tarsilla, 2014), while evaluation aspires to play a strategic learning function. This chapter explores this tension by looking at the extent to which evaluations in the region are answering strategic questions, versus monitoring processes and outputs. This is important because there is a widespread assumption that a spread of evaluation systems both reflects and supports a concomitant expansion of a culture of learning and evidence-based decision-making. It is possible, however, that if there is not a shared view of the organisational role evaluations play, the evaluation function will simply become another tool of compliance.

Practitioners and commissioners often differ on the purpose for the evaluation function and, as a result, they place it differently on the spectrum from management to governance. Since evaluations also have a disciplinary focus based on the programme being evaluated, such as education, health and agriculture, the more subtle difference of organisational location often goes unnoticed. Different stakeholders in the evaluation process often make tacit assumptions about which level on the results chain an evaluation is targeting for decision-making, as well as appropriate evaluation questions and methodologies.

These differing assumptions often manifest as tension around the most appropriate process, method and product for an evaluation.

This is not a challenge that is new to evaluation. One of the founding fathers of the field described evaluation as “an overdue child of allegedly incompatible parents whose families opposed the marriage. And its difficult birth began a confused childhood exhibiting multiple personality disorder” (Scriven, 1994). One of the results of this disorder is that the diverse stakeholders of an evaluation process, who often hail from diverse disciplinary backgrounds themselves, often fail to establish shared aims of an evaluation process.

Furthermore, they may not fully understand and consider differences in values and perspectives, the appropriateness and utility of an evaluation (Reeve & Paperboy, 2007). Other authors frame this issue as a “lack of unified approach” (Khan, 1998).

One of the key fault lines of this debate is whether evaluation sits primarily in the field of management, or governance. This debate is as old as the practice of evaluation itself. Linimon and Borish (1981) describe a continuum in this role, “from the evaluator as a technician, to the evaluator as a change agent. At one pole of the continuum, the evaluator is portrayed as an illuminator, fulfilling the technical functions of information gathering and reporting for the party commissioning the evaluation. As the evaluator role shifts from this pole, the evaluator fulfils the additional functions of interpreting information, describing the social-political context of the evaluation, and identifying and advocating alternative courses of action.” Often in evaluation reports, it is possible to tease out differences of opinion held by the commissioning organisation and the evaluator; evaluators often aim to place the evaluation function in a more strategic governing role, while commissioning agencies are more likely to locate it as a managerial requirement, with a tighter focus on outputs and compliance.

While the authors do not hold a normative view around a single, optimal way to locate the evaluation function, an important foundation to understanding the evaluation landscape in Africa is understanding the nature of current practice. This chapter explores the purpose of evaluations in the database, to try and understand whether evaluations are primarily sitting in the domain of management, or evaluation. Evaluations performing a management function are fulfilling monitoring objectives, of determining the extent to which predetermined goals and objectives have been achieved. Evaluation in the governance domain has a more strategic purpose, namely to understand the way in which interventions have worked in context, consider alternatives and explore whether the objectives set have been the right ones.

It is particularly important to understand this distinction, because one of the critiques of the monitoring and evaluation field in Africa is that it is donor-led and practice is still driven by non-Africans (Khan, 1998; Porter & Goldman, 2013). First, while the problem is widely acknowledged, it is usually mentioned as an assertion made without empirical backing. The AfrED database provides an opportunity to begin bringing evidence to this debate. The suggestion is that

donors may come with a different approach to the evaluation function, different methods of measurement, as well as different definitions for success. As a result, evaluations may not be contextually relevant (Kabore, 2013; Khan, 1998). However, data on this is limited, and this chapter will make some contribution towards this gap, interrogating whether various commissioning agencies place evaluation as a management or a governance function.

One more point of contestation in discussions of the African evaluation landscape is on the balance between monitoring and evaluation (Cloete, Rabie & De Coning, 2014). It is often assumed that, as monitoring and evaluation systems expand, monitoring is privileged over evaluation. Sufficient investment in systems of evaluation is rare. While monitoring is critical to build effective systems of management and understand processes of implementation, it is insufficient to make decisions of programmatic strategy and design. While evaluation has been the centre of attention over the last decade, it has long been recognised that even systems using the word evaluation may not necessarily fulfil an evaluative function. As one scholar pointed out nearly thirty years ago, "In fact, there was a wide gulf between evaluation precepts and practice. In practice, most evaluative reports of many donor agencies were monitoring types of efforts, concentrating on implementation issues rather than on assessing development results" (Binnendijk, 1989). Unfortunately, the debate does not seem to have shifted significantly in the intervening years, perhaps in part due to the dearth of empirical data to propel the debate.

While public sector M&E systems in the region put more emphasis on monitoring than on evaluation (Cloete et al., 2014; Abrahams, 2015), one of the findings of this study is that there are still several layers to peel away before we can reach any conclusions about the balance of monitoring and evaluation, and their implications for both evaluation practice and evaluation demand and use.

Finally, questions remain about the role of evaluations within the current public sector culture (Goldman, Mathe, Jacob, Hercules, Amisi, Buthelezi, Narsee, Ntakumba & Sadan, 2015; Cronin & Sadan, 2015). By looking at the purpose of evaluations as they currently stand, it will become more apparent whether or not evaluation can be a tool to either trigger changes in public sector culture and norms, or, alternatively, better understand the limitations of evaluation to catalyse change (Mayne, 2017).

Background to management and governance in Africa

The lines between management and governance are not necessarily clear cut; with management being the implementation system of governance, the two functions can be described colloquially as “rowing” versus “steering”. Governance explains how goals are set, while management determines the processes through which they are achieved.

Monitoring and evaluation in Africa have historically been linked to the donor space. One of the implications of this is that monitoring and evaluation related reforms have often been introduced to the public service by institutions of Treasury, linking the function to audit and compliance (Abrahams, 2015). The compliance-oriented nature of state practice has had a range of implications for the development of the monitoring and evaluation sector. One is that there is a limited choice of methods and analysis available to evaluators. Variable data systems have made evaluation practice costly and make the quality of M&E systems and practice variable.

From a governance perspective, this association with the donor space has created questions about how decisions are made and, in evaluation design, who decides on the most appropriate approach. There are questions of whether donor-led evaluations are contextually relevant, and whether there are appropriate power relationships between commissioners and users of evaluation. Finally, there is a risk that the learning loop is broken, given the different stakeholder influences.

Leaders within the evaluation sector often share an expectation that evaluation does not exist “merely to assess merit or worth in any given programme or project, but also to establish and institutionalise roles, relationships and practices among a wide breadth of development partners” (Rutkowski & Sparks, 2014) and “a component of political strategy” (Taylor, 2005). These expectations may not be shared by evaluation commissioners. This chapter explores the extent to which the evaluations actually carried out primarily address issues of management, or of governance. It begins a conversation about how different stakeholders are negotiating the role of evaluation, its implications for evaluation methods and ethical practice.

Research methods

The objective of the study on which this chapter is based is to use the evaluations in the AfrED database to develop an empirical foundation to better understand the current purpose of evaluations being undertaken in Africa. This may be reflected in patterns identified in the purpose section of evaluations and methods employed, or trends linked to sectoral areas of work or commissioning agencies. The hope is that by mapping some empirical results, we will uncover additional questions around the current purpose of evaluations. In a context where national evaluation systems are growing within states that often focus heavily on compliance, it is important to understand if evaluation systems are 'pulling' states into a learning space, or, alternatively, if evaluations are adapting to a monitoring function.

In order to do this, the first step was to map out an approach that would allow us to categorise the evaluations, and then to identify emergent patterns.

Inspired by systematic reviews

In order to locate the evaluations in the AfrED database in the fields of either governance or management, this study drew methodological inspiration from a systematic review approach. The study did not carry out a full systematic review, nor would it have been appropriate, given the database already providing a limited population of evaluations. However, following the steps of developing, testing, refining and implementing a protocol to categorise evaluations as playing a management role, a governance role, or both, drove the research approach. The study aimed to develop a tool that would allow for a replicable process for evaluations outside of the database population. With this in mind, the first five of eight steps outlined in the Cochrane Handbook (Higgins & Green, 2005) were adapted for this specific study as follows:

1. *Defining the review protocol, and developing criteria for including evaluations*

A first hurdle in this study was creating categories that broadly described the purpose of the evaluation exercise and its implications for both the location of the field from a professional standpoint, and the level of the results chain from a commissioning and use standpoint. To respond to these multiple goals, we found that the broad categories of management and governance best

encapsulated the concepts we were trying to apply. They both matched to disciplines in higher education and broad purposes of evaluation – from the commissioning and use side.

While there are some areas of overlap between management and governance, they are concepts that are widely understood as fulfilling different functions within the organisation. With a bit of iterative research, we could find mutually exclusive definitions that would serve the purpose of the study. Building on certain consensus definitions (Fukuyama, 2013; Dayton, 2001; UNESCO, 1998; De Janvery & Dethier, 2012), the following components were identified to allow us to describe the categories of management and governance more comprehensively.

Table 2.1 Identifying the core constructs of management and governance in evaluations

Management	Governance
Efficiency	Setting strategy
Working to a goal	Stakeholders
Implementing	Situational analysis/ context
Meeting targets	Relevance
Adhering to norms, standards and policies	Setting a goal
	Setting norms, standards and policies

These constructs were then developed into a protocol which would be used to categorise evaluations as sitting primarily in either the discipline of management or in the discipline of governance, or broadly across both disciplines.

2. *Selecting evaluations*

While a true systematic review would have cast a wide net to identify evaluations, this study simply sampled within the database to search for evaluations. In our initial plan to sample the database, we anticipated that some subset of evaluations would explicitly discuss management or governance. We hoped to construct a purposive sample through searching for the terms directly. An initial sample of the database using the search terms of management and governance very quickly demonstrated that this was not the case. An unanticipated challenge of terminology was that a large number of the evaluations in AfrED database are in the health sector, and of these many deal with the management of disease.

While rigorous definitions of concepts of management and governance are critical to carrying out the review, this challenge demonstrated very early on that no shortcut to conceptualising these two areas would be possible.

Realising that the initial approach would not yield the expected results, a random sampling strategy was chosen. This would allow the authors to analyse results across the broadest range of sectors and countries, and also help uncover trends across the database as a whole. Furthermore, by constructing a representative sample, the authors hoped to refine the protocol in such a way that the conceptual constructs would be theoretically sound across differences in commissioning agency terminology or thematic area of work.

This study used Excel as a random sampling statistical package. Excel allocated a random number adjacent to each evaluation name in a separate column in a spreadsheet. Once each item was assigned a random number, the spreadsheet was ordered from the smallest random number to the largest. In order to arrive at an appropriate sampling size, a power analysis was conducted. The sample size is based on a 95% level of statistical confidence. From a population of 2635 evaluation reports in the AfrED database, a sample size of 163 was achieved.

3. Testing and refining the protocol

Once the sample was constructed, the protocol itself was tested and refined. To do this, an additional ten evaluations were brought in as a trial, and three reviewers applied the protocol to the same ten evaluations to see if it was valid. It was revised and retested twice with additional sets of ten evaluations before the researchers had a construct for each discipline that was sufficiently valid. Through the testing process, two additional categories were created. One was both management and governance, for evaluations which clearly included at least one component from both categories. A fourth category was created for evaluations that were primarily focused on methods, or the evaluation process itself, since these evaluations did not fit into any of the categories created. They were nevertheless included in the AfrED database.

As part of the process of trailing the protocol, the research team discovered that few evaluations were explicit about both the purpose of the evaluation, and the way in which their methods achieved this purpose. This reflected previously discussed differences in perspective between evaluation commissioners and

practitioners. Rather than being a quick process of matching stated objectives to the categories developed, categorising the evaluations was a much more analytical process of teasing out and clarifying the stated objectives, interpreting them (based on what was actually done in the evaluation process), and then looking at the level of results and conclusion.

4. *Applying the protocol*

Due to the challenges above in constructing a valid protocol, the researchers chose a rigorous process of applying the protocol. Two senior researchers categorised each of the 163 evaluations into one of the four categories (management, governance, both, or methods) indicating at least two components of the category from Table 2.1. The researchers compared lists and identified all differences. Each difference was individually discussed and resolved, either as an issue of human error or substantive disagreement. In the cases of human error, the mistakes were rectified, and the evaluation was considered categorised. In the cases of substantive disagreement, each researcher presented their logic; in most cases, this process identified an area of either substantive oversight or logical fallacy. This was then rectified and the evaluation categorised. In two of the 163 cases the researchers held substantive disagreements and a third senior researcher was called in to moderate. Through this process, consensus was reached on the categorisation of each evaluation in the sample. Throughout the process, researchers kept detailed reflective journals of lessons or questions that emerged through the categorisation process.

Once all evaluations in the sample were categorised, data was analysed according to sector, commissioning agency, and country. Additionally, a thematic analysis of the protocol implementation journals informs the structure of the results presented below.

Methodological challenges

This research provides a strong empirical basis to argue that evaluations currently undertaken in Africa are overwhelmingly managerial in nature. They serve more of an output monitoring function than a platform for strategic decision-making. However, this piece of research is far from exhaustive and faces a number of limitations that are worth noting.

The construction of the database itself

The AfrED database was quite exhaustive in its collection of academic research articles on monitoring and evaluation. However, due to the volume and scope of evaluations undertaken in the region, only a subsection of evaluations conducted during the period covered by the database were able to be included. This will be built in the coming years. The database thus does not represent the full population of evaluations undertaken in the region, so the results of this study remain indicative rather than rigorously generalisable. There are important commissioning agencies that are not reflected in the database; they may bring different approaches.

What is an evaluation?

One unexpected challenge of this study was determining what exactly constitutes an evaluation. While this was dealt with in detail in the methodology chapter for the purposes of database construction, we confronted a number of situations where the database included studies that were not traditional programme evaluation reports. Some examples of these include:

- ▶ A national evaluation policy framework;
- ▶ A barometer on the national development assistance environment;
- ▶ A process document on the development of a monitoring and evaluation system;
- ▶ Case studies from evaluative processes; and
- ▶ A national level sectoral study.

While all of the above are clearly linked to evaluation in a range of ways and some stem from one or more specific programme evaluations, the products of evaluative processes are not necessarily at the same level. Documents that were not traditional evaluation reports were at times difficult to assess using tools that were designed for evaluation reports only. This did not represent a large portion of the sample, but did make it apparent that while certain organisations have standardised evaluation report structures, in many cases there is a spectrum of products – from research reports to country synthesis studies to barometers. Drawing the line around what exactly is an evaluation was not always a straightforward process.

Developing a shared conceptual approach to evaluation purpose

One of the most difficult methodological challenges in categorising the evaluation reports was that terminology varied widely across evaluations. There was very little standardisation even within an evaluation report. The stated aims of the evaluation were often not, upon examination, what was done. While this lack of standardisation is common in an emergent profession like monitoring and evaluation, this may be symptomatic of the gap between evaluation supply and demand (Stem, Margoluis, Salafsky & Brown, 2005). Commissioners, who define the purpose of an evaluation in terms of reference, may not successfully develop a shared vision with the evaluation consultant who carries out the evaluation. This results in a disjuncture between the stated aims and the actual work carried out.

Results

This study clearly demonstrates that across a range of fault lines there are very different views on both the purpose and practice of evaluation. Commissioning agencies differ in their approach to the evaluation function; organisations locate evaluations at different levels of the results chain, and design evaluations for different purposes. Furthermore, different sectors have different norms around evaluation practice and purpose. Health and education, two sectors with a large number of evaluations in the database, seem to target evaluations lower on the results chain than other sectors. Finally, the purpose of evaluations differs by country as well. Interestingly, countries that are leading in evaluation systems development are those that are also targeting their evaluations lower down on the results chain. As country led evaluation systems grow in Africa, understanding these differences and their implications for the evaluation sector, will be important to building systems that are useful and contextually relevant.

Description of the evaluations

For the purposes of this study, 163 evaluation reports were randomly selected from the database. Figure 2.1 describes the sample by sector. Most of the evaluation reports in the sample are public health focused. This is indeed a true reflection of the AfrED database as public health represents the biggest group of evaluations in the database. Evaluations focusing on governance and formal

education also account for more than 10% of sample proportions (15% and 10% respectively). Health, public policy (including donor strategy) and education are amongst the highest areas benefiting from evaluation studies across the region. Other significant sectors include clinical health – mainly driven by the level of resourcing devoted to the Global Fund on Malaria, Tuberculosis, and HIV/AIDS, in the past decade (World Health Organization, 2014); social cohesion and poverty reduction, which are both mainly driven by high migration levels in the region (UN Department of Economic and Social Affairs, 2016).

Figure 2.1 Evaluations by sector (based on AfrED database)

Figure 2.2 shows a breakdown of sample evaluations by evaluating agency. Considering the list of agencies and their respective work in social development in Africa, one can speculate about the purpose of the respective evaluations in the sample. This is especially the case with an organisation such as 3ie which funds many impact evaluations and generally promotes highly scientific evaluation methods for programme effectiveness. The sample also shows evaluation leadership by international agencies broadly. Finally, it is encouraging to observe a relatively significant representation of evaluations received from country departments and various spheres of government (at 14.3% of the sample).

Figure 2.2 Evaluations by agency (based on AfrED database)

A common assertion in the field of monitoring and evaluation is that the field in Africa remains donor led, and that this comes with a range of related problems of appropriateness, incentive structures, and ownership (Bornstein, 2006). While the initial construction of the database does not yet allow us to make conclusive empirical findings on this, it does provide enough information to begin a conversation that is grounded in research. Figure 2.2 does suggest that this assertion holds merit. Fewer than one in six evaluations in the database are commissioned by a national agency, with all the rest led by bilateral or multilateral donors. Research is still emergent on the ways in which commissioning organisations shape evaluations (Porter & Goldman, 2013). There is a strong push globally to strengthen national ownership of evaluations and evaluation systems (Green, 1988; Kirby et al., 2003).

Figure 2.3 Evaluations by country (based on AfrED database)

Figure 2.3 shows a breakdown of evaluation by evaluation country. Three clear trends emerge from this analysis:

1. There is a growth in popularity of multi-country programmes and programme evaluation. Perhaps this can be attributed to the emerging popularity of contextual studies in evaluation practice, especially by donor agencies.
2. The obvious interest of the donor community in Kenya and Ethiopia as popular destinations for donor funding. These countries have also featured significantly in multi-country evaluations.
3. Finally, it is also interesting to observe evaluation production rates of South Africa and Uganda as the two countries with the most mature national evaluation systems (NES) in the region (Genesis Analytic, 2017). The sample shows that most of the evaluations from the two countries are government commissioned as opposed to donor commissioned as is the case in other countries.

The purpose of evaluation

Of 163 evaluations, more than half exclusively measure progress towards predetermined results, while only 12% of evaluations have an exclusive focus on the strategic positioning of programmes. This strongly endorses the hypothesis that evaluation practice in the region emphasises progress towards specific results, rather than looking more broadly at programme strategy in context.

Given the existing research indicating that monitoring and evaluation practice in Africa heavily emphasises monitoring (Cloete, 2009), it is important to highlight that all evaluation is not equal. If evaluations are functionally monitoring implementation processes, then even emergent evaluation practice is still not serving the needs of an evaluation system but is rather an extension of the purpose of monitoring.

Figure 2.4 Overall purpose of evaluation (based on AfrED database)

This research does not come with a normative prescription on the ideal balance of management and governance in evaluation practice. However, given the criticism that the field of monitoring and evaluation is not enabling of strategic decision-making, due to a strong compliance focus (Govender, 2013), some sort of balance needs to be struck. At this moment, empirical evidence from the database demonstrates a disproportionate focus on monitoring functions. Clearly there is a need to shift the evaluation function into a more strategic decision-making position.

Looking at the purpose of evaluation by sectoral area, there is variation across sectors. Over 80% of all evaluations in the education sector is in the realm of management, versus less than half that percentage of evaluations with a broad social focus. On one hand, it seems appropriate that health and education lend themselves better to quantitative approaches. They are measuring progress on agreed upon outcomes, since these established fields have significant

collective agreement around pre-determined outcomes. More emergent fields appropriately take more exploratory approaches. However, this may also illustrate a limitation in the ability to innovate in the fields of health and education.

Figure 2.5 Purpose of evaluation by sector (based on AfrED database)

One important motivation for this study is to understand the current purpose of evaluation in the region in the context of a growing demand for country-led evaluation systems. This is an imperative of the Sustainable Development Goals regime. Progress in developing national evaluation systems in Africa has been uneven, but momentum is growing rapidly (Goldman et al., 2018). As the national imperative for strengthening evaluation systems grows, it is important to understand the variation of evaluation purpose amongst countries in the region.

Figure 2.6 illustrates evaluation purpose by country within the database. It demonstrates two important trends. The first is that South Africa and Uganda, the two countries with the strongest evaluation systems in the region, demonstrate the least strategic evaluation purpose. As the progress continues towards increasing country ownership of evaluation systems, this finding indicates that country ownership in itself is not sufficient for a robust evaluation system. It is but one component of a successful evaluation system, which should include a consideration of the purpose and results of evaluations themselves.

Figure 2.6 Evaluation purpose by country (based on AfrED database)

A second interesting result, reflected in Figure 2.6, is that multi-country studies are significantly more likely to deal with more emergent issues and use evaluations for strategic decision-making and programme governance. More research is needed to confirm why this is the case, but possible hypotheses could include the need for relatively more groundwork to understand the context across multiple countries. Alternatively, there may be something inherent in the nature of multi-country programmes, whether it is in their funding or institutional arrangements, that makes resources more likely to be available for an evaluation that addresses issues of programme governance.

It is clear from the discussion above that currently evaluation is primarily playing a management function in the region. However, there are notable exceptions, with specific evaluations, commissioning agencies, and sectors which tend to move more into the governance space. It is worthwhile to look at the pockets of innovation that can be found. From our samples, Kenya is doing the most evaluations in the governance space, and this is happening across a range of thematic areas. In this case, public health is an important driver of innovation (as opposed to clinical health research, which is largely less exploratory). While education does not feature, early childhood development does. Evaluations with a broad social remit, such as from fields of poverty reduction and governance, tend to maintain a strategic focus.

Figure 2.7 Governance-focused evaluations in Kenya (based on AfrED database)

Fit-for-purpose methods

The section above clearly demonstrates the lack of consensus amongst various stakeholders about the purpose and scope of the evaluation function. There are variations across organisations and also differing norms in different disciplines. When data collection began, there was an intention to collect data on the methods employed for each evaluation in the sample. However, in keeping with the findings of other authors in this volume, such as in Chapters 4 and 5, evaluations do not often explicitly discuss the methods used, so this approach was dismissed. Rather, methodological findings and trends were noted in the researcher journals, and these were drawn on for discussion below.

Figure 2.8 demonstrates that different organisations have different understandings of the purpose and practice of evaluations. As the authors of Chapter 5 in this volume highlight, this differing understanding of the evaluation functions also translates into different methods employed through evaluation practice. Some of this variation can be understood by considering the nature

of the organisation in question. It seems appropriate that ALNAP, which deals with humanitarian evaluations, should focus on emergent practice and look extensively at strategy in context.

Figure 2.8 Evaluation purpose by commissioning agency (based on AfrED database)

3ie, on the other hand, is an organisation dedicated primarily to impact evaluations. ‘Impact’, like other terms in the evaluation sector, is used differently by different stakeholders, creating some confusion around semantics (White, 2010). While the organisation takes a theory-based approach to understanding programmatic contribution to certain results, the organisation also defines impact evaluation as a specific quantitative methodological approach to understanding the attribution of a specific indicator. This single evaluation method answers critical research questions about programmatic contribution, but is explicitly not well equipped to consider questions of context, alternatives or unintended consequences. This is well understood by 3ie and other experienced evaluators, which is why other approaches to impact evaluation are also available (Rogers, 2009), and this specific method is considered one tool in the arsenal of evaluators. However, as a result, nearly every evaluation led by 3ie is measuring progress towards predetermined goals and objectives. There has been a tremendous growth in both the resourcing and

popularity of quantitative, RCT-based impact evaluations of social programmes over the last decade (White, 2010). This is an important shift to recognise and understand in the evaluation landscape, and consider what it means for the role evaluation plays in organisational practice and development.

This raises one area for exploration in fit-for-purpose evaluation methods. There is a large body of research in the field of methods that suggests people 'trust' results stemming from quantitative research more, and that qualitative research is more strongly tied to trust in the researcher as an individual (Kuhn, 2012; Patton, 2002). While there is also ample research presenting a counterpoint around qualitative research approaches, one particular challenge is that with a utilisation-focused approach to evaluation, perceptions may be more critical than validity of data to ensuring ownership and use of resultant evaluations. This study does not presume to enter into a debate on a hierarchy of research methods. However, an important area for future research is to understand the African context around credibility of methods, appropriateness of these methods (which Tirivanhu and Blaser Mapitsa in this volume explore in further detail, see Chapter 7). Empirically grounding this debate may help illustrate some of the challenges evaluation practitioners face in meeting the aspirations of an evaluation that is of strategic importance, with 'credible' methods and approaches.

Similarly, the OECD takes a prescribed approach to evaluations, adhering strictly to the OECD/DAC criteria of relevance, effectiveness, efficiency, impact and sustainability. These criteria have a long history in the evaluation space, but have recently been subject to discussion on areas of improvement and their ongoing contribution to the evaluation space (Chianca, 2008). Comparing these criteria to the components of the protocol originally developed on management and governance, some of the criteria fall into the realm of governance, while others fall into the realm of management. This suggests that most OECD/DAC oriented evaluations should be categorised as fulfilling both a management and governance purpose. However, digging deeper into the way the criteria are applied in practice demonstrated that this evaluation approach was applied narrowly in most cases, looking only at programme implementation and achievement of immediate results.

Supply and demand gaps reflected in the landscape

One of the most striking findings of applying the protocol to the evaluations in the sample is that very often there was significant discrepancy between the purpose as stated in the introduction of the evaluation, and the actual methods and results in the evaluation undertaken. In over half of the evaluations which were not categorised in the same way by researchers in the first round of applying the protocol, it was because the researchers primarily drew on evidence from different sections of the evaluation. Most often, the introductory section which stated the purpose of the evaluation placed the evaluation purpose firmly in the strategic governance space. However, after looking at what was actually done, it was clear that the evaluation could only serve a managerial purpose. This suggests that there is an aspiration by some stakeholders for evaluation to play a governance function, but there are barriers in place in the evaluation process. Future research may need to focus on better understanding the nature of these barriers.

On the demand side, the construction of public sector evaluation systems is such an important feature of the evaluation landscape in Africa at the moment, a closer look at evaluations commissioned by the public sector is useful. While we saw above that, in general, evaluations commissioned by government do not tend to be exploratory in nature, Figure 2.9 demonstrates that this aggregate analysis is not sufficient, and that there is considerable variation in the purpose of evaluations within government. First, the figure reinforces the finding above, that non-state evaluations are better suited for innovative learning. Local government in particular is conducting evaluations that are concentrated lower on the results chain. While further work is needed to explore this, it could be a reflection of the compliance-driven nature of much local government in the region or could demonstrate a reliance on monitoring systems for evaluation practice. Finally, it is interesting to note that within the public sector, evaluations in the education space are more exploratory than evaluations in the education sector as a whole, as reflected in the database. This could be linked to methods employed, or it could indicate that governments in the region have an urgent need to learn and innovate in the education sector.

Figure 2.9 Evaluation purpose by public sector ministry
(based on AfrED database)

On the supply side, there is limited empirical data looking at the identity of evaluators in Africa, and perceptions around the credibility of evaluation results and subsequent evaluation use. However, one of the key assertions in the regional evaluation debate is that non-African evaluators do not come with sufficient contextual understanding of the region to generate appropriate and useful evaluations. This is being explored in more detail by Tirivanhu and Blaser Mapitsa (see Chapter 7), but institutional affiliation of individuals is also important to consider. Internal evaluations are one clear manifestation of an approach to bridge the supply and demand gap, but they have been insufficiently interrogated in the African evaluation literature. In the AfrED database, over a quarter of evaluators report an internal affiliation, and nearly a third were some combination of internal and external. Understanding how this shapes the purpose of evaluations and concomitant methods will be an important next step.

Figure 2.10 Institutional local of evaluator

Global literature offers some information about the institutional identity of evaluators and perceptions of credibility. Even though internal evaluations are often of equal or superior quality to externally commissioned evaluations, those conducted by someone external to the organisation are usually considered more credible (Conley-Tyler, 2005). Anecdotal information suggests this may also be the case in Africa and raises additional questions about whether there is an outsider-credibility bias. Given the predominance of questions around donor-linked evaluation practitioners in the region (explored by Ngwabi in this volume, see Chapter 4), it is an important area for further exploration.

Additional research is needed to understand these phenomena in detail, but the supply and demand gap in the region reflects a combination of factors, including the differing perspectives of stakeholders on the purpose and scope of evaluations, as well as variation in resources available for the evaluation process. On the one hand, there are some incentives for some evaluation practitioners, particularly those who also identify as researchers, to keep evaluations high on the results chain. At the same time, there are incentives to reduce the scope of work for contracted consultants and keep the evaluation remit narrow and specific. It will be important to better understand how different stakeholders articulate these divergent incentives, but from the data in this study it is apparent that the stated objective is more strategic than the methods as implemented.

When evaluators and authors of evaluation related research were surveyed in the African database about the purpose of the evaluations they conducted, the data presents a dilemma. The most important finding is that accounting to donors is the least common purpose of evaluations, with 109 of 460 respondents never conducting evaluations for accountability. Decision-making and improving practice are the most common purposes. However, organisational learning also remains relatively insignificant, with only 179 respondents frequently conducting evaluations for learning. This may suggest that this gap between intention and application of the evaluation function is a manifestation of the gap between supply and demand.

Figure 2.11 Purpose of evaluations conducted (AfrED database close out report)

What these anecdotal experiences underscored is that there is little consensus about the purpose of monitoring and evaluation, and where on the results chain or organisational practice the monitoring and evaluation function sits. So far, the thrust of the debate has been around professionalisation of evaluation (Basheka & Byamugisha, 2015), taking an angle of evaluation practice. However, the evaluations on the AfrED database provide us with an opportunity to look at the evaluations themselves, and see what the stated purpose of the evaluation was, whether the methods were appropriate to the purpose, and to what extent, from the commissioning side, it appears that evaluations are being undertaken to address results at a low level of the results chain, or a high

level of the results chain. As different role players in the region are exploring the gap between the supply and demand for evaluation (Abrahams, 2015), we hope that this chapter will help define this gap by identifying key questions around the stated purpose for evaluations, whether the methods and evaluation implementation responded appropriately to this purpose, and what implications this may have for both the professionalisation debate and the discussions on evaluation demand and use.

Conclusion

The empirical results from the database debunk some myths about evaluation purpose and practice in Africa, but confirm truth in others. The primary result of this study is finding that evaluations in the region have a disproportionate focus on the low end of the results chain, focusing on processes of implementation and progress towards predetermined results, rather than broader analyses triangulating context, implementation process, and results. This means that even though evaluation practice in the region is growing, the purpose of evaluation is still very closely tied to monitoring. This both reflects a lost opportunity for learning, but also calls into question whether an expansion of evaluation systems in the region is likely to achieve the transformative affects on culture and programme design that are anticipated. Several core purposes of the evaluation function, such as considering alternative programmes to achieve the same objective and passing judgement on the choice made, are not reflected in current evaluation practice.

Linked to the finding above, is a demonstration that various stakeholders, from evaluation practitioners to commissioners, do not necessarily share a view on the purpose of evaluation, and do not necessarily hold common definitions of the distinction between monitoring, and evaluation. As Jane Davidson, one of the founders of evaluation, said, "the lack of a clear distinction between the two means that evaluation is getting short-changed" (Davidson, 2013). This is certainly the case in the evaluations studied.

The growth of evaluation systems in the region present a unique opportunity to begin to build this consensus around the purpose and practice of evaluation in the region. However, this study has demonstrated that public-sector led evaluations in particular marginalise the evaluatory function, and are instead using evaluation as an extension of monitoring and compliance. As the role

governments have to play in building evaluation systems grows, it is important both to be aware of this dominant ideology, and its implications for the growth of the evaluation sector. It is particularly important to continue the currently dynamic discussion around what the role of evaluation is, and should be, in the context of Africa's public sector management systems.

An additional finding this chapter explored, but was not able to conclude, is the relationship between the purpose of evaluations, and the methods employed. It is appropriate that evaluations that were centered around a large quantitative component focused on generalisability, while evaluations that were looking more at emergent context employed qualitative methods. However, the strength of this linkage, in addition to other components of evaluation methods, such as degrees of participation, and the role of the commissioning organisation, are an important next step in further understanding the purpose and practice of evaluation in the region. This is a particularly important area of exploration given the current global trends towards impact evaluations, randomised control trials, and other methods that may not be best suited to responding to questions of context. Evaluation systems that grow around an assumption that these methods represent a 'gold standard' will, in all likelihood, be seen as a more managerial tool.

There is a need for greater exploration amongst all stakeholders around the purpose of evaluations in Africa, and for the discourse both professionalisation of the sector, contextually appropriate evaluation methods, and purpose of evaluation systems in the region to be grounded in empirical data on current practice. Furthermore, this foundation of data should provide a focal point for stakeholders with diverse views to engage, and, if not reach areas of consensus, at least acknowledge areas of disagreement to minimise tension. At this moment, accusations of poor-quality evaluations, inappropriate methods or unskilled evaluators are often masking a fundamental disagreement around the purpose and potential of evaluations.

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>
- Bamberger, M.; Rao, V. & Woolcock, M. 2010. *Using mixed methods in monitoring and evaluation: Experiences from international development*. Washington, DC: The World Bank.
- Basheka, B.C. & Byamugisha, A. 2015. The state of Monitoring and Evaluation (M&E) as a discipline in Africa. *African Journal of Public Affairs*, 8(3):75-95. <https://doi.org/10.4102/apsdpr.v3i2.80>
- Binnendijk, A.L. 1989. Donor agency experience with the monitoring and evaluation of development projects. *Evaluation Review*, 13(3):206-222. <https://doi.org/10.1177/0193841X8901300302>
- Bornstein, L. 2006. Systems of accountability, webs of deceit? Monitoring and evaluation in South African NGOs. *Development*, 49(2):52-61. <https://doi.org/10.1057/palgrave.development.1100261>
- Chianca, T. 2008. The OECD/DAC criteria for international development evaluations: An assessment and ideas for improvement. *Journal of Multidisciplinary Evaluation*, 5(9):41-51.
- Cloete, F. 2009. Evidence-based policy analysis in South Africa: Critical assessment of the emerging government-wide monitoring and evaluation system. *Journal of Public Administration*, 44(2):293-311.
- Cloete, F.; Rabie, B. & De Coning, C. (eds). 2014. *Evaluation Management in South Africa and Africa*. Stellenbosch: African Sun Media. <https://doi.org/10.18820/9781920689513>
- Conley-Tyler, M. 2005. A fundamental choice: Internal or external evaluation? *Evaluation Journal of Australasia*, 4(1/2):3. <https://doi.org/10.1177/1035719X05004001-202>
- Cronin, G.P. & Sadan, M. 2015. Use of evidence in policy making in South Africa: An exploratory study of attitudes of senior government officials. *African Evaluation Journal*, 3(1):10. <https://doi.org/10.4102/aej.v3i1.145>
- Davidson, J. 2013. Monitoring and evaluation: Let's get crystal clear on the difference. *Genuine Evaluation*. <http://genuineevaluation.com/monitoring-and-evaluation-lets-get-crystal-clear-on-the-difference/>
- Dayton, N.A. 2001. Total quality management critical success factors, a comparison: The UK versus the USA. *Total Quality Management*, 12(3):293-298. <https://doi.org/10.1080/09544120120034447>
- De Janvry, A. & Dethier, J.J. 2012. *The World Bank and governance: the Bank's efforts to help developing countries build state capacity*. The World Bank. <https://doi.org/10.1596/1813-9450-6275>
- Fukuyama, F. 2013. Democracy and the Quality of the State. *Journal of Democracy*, 24(4):5-16. <https://doi.org/10.1353/jod.2013.0074>
- Higgins, J.P. & Green, S. 2005. *Cochrane handbook for systematic reviews of interventions*. Hoboken, NJ: Wiley-Blackwell.
- Goldman, I.; Byamugisha, A.; Gounou, A.; Smith, L.R.; Ntakumba, S.; Lubanga, T.; Sossou, D. & Rot-Munstermann, K. 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>

- Goldman, I.; Mathe, J.E.; Jacob, C.; Hercules, A.; Amisi, M.; Buthelezi, T., Narsee, H.; Ntakumba, S. & Sadan, M. 2015. Developing South Africa's national evaluation policy and system: First lessons learned. *African Evaluation Journal*, 3(1):1-9. <https://doi.org/10.4102/aej.v3i1.107>
- Govender, I.G. 2013. Monitoring and evaluation systems for enhancing governance and government legitimacy in South Africa. *Journal of Public Administration*, 48(Special issue 1):811-823.
- Jacob, S. & Boisvert, Y. 2010. To be or not to be a profession: Pros, cons and challenges for evaluation. *Evaluation*, 16(4):349-369. <https://doi.org/10.1177/1356389010380001>
- Jesover, F. & Kirkpatrick, G. 2005. The revised OECD principles of corporate governance and their relevance to non-OECD countries. *Corporate Governance: An International Review*, 13(2):127-136. <https://doi.org/10.1111/j.1467-8683.2005.00412.x>
- Kabore, I. 2013. Commentary on country-led monitoring and evaluation systems, better evidence, better policy, better development results. *African Evaluation Journal*, 1(1):1-3. <https://doi.org/10.4102/aej.v1i1.46>
- Khan, M.A. 1998. Evaluation capacity building: An overview of current status, issues and options. *Evaluation*, 4(3):310-328. <https://doi.org/10.1177/13563899822208626>
- Kirby, P.; Lanyon, C.; Cronin, K. & Sinclair, R. 2003. *Building a culture of participation. Involving children and young people in policy, service planning, delivery and evaluation*. London: National Children's Bureau.
- Kuhn, T.S. 2012. *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Linimon, D. & Borich, G. 1981. Evaluator as technician versus evaluator as change agent: A continuum of power relationships. Paper presented at the Joint Annual Meetings of the Evaluation Network and the Evaluation Research Society, Austin, Texas.
- Marshall, M.N. 1996. Sampling for qualitative research. *Family Practice*, 13(6):522-526. <https://doi.org/10.1093/fampra/13.6.522>
- Mathison, S. 1994. Rethinking the evaluator role: Partnerships between organizations and evaluators. *Evaluation and Program Planning*, 17(3):299-304. [https://doi.org/10.1016/0149-7189\(94\)90009-4](https://doi.org/10.1016/0149-7189(94)90009-4)
- Mayne, J.W. 2017. *Monitoring performance in the public sector: Future directions from international experience*. New York: Routledge.
- Mouton, C. 2010. The history of programme evaluation in South Africa. PhD thesis, Stellenbosch University.
- Patton, M.Q. 2002. Two decades of developments in qualitative inquiry: A personal, experiential perspective. *Qualitative Social Work*, 1(3):261-283. <https://doi.org/10.1177/1473325002001003636>
- Porter, S. & Goldman, I. 2013. A growing demand for monitoring and evaluation in Africa. *African Evaluation Journal*, 1(1):1-9. <https://doi.org/10.4102/aej.v1i1.25>
- Reeve, J. & Paperboy, D. 2007. Evaluating the evaluation: Understanding the utility and limitations of evaluation as a tool for organizational learning. *Health Education Journal*, 66(2):120-131. <https://doi.org/10.1177/0017896907076750>
- Rogers, P.J. 2009. Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation. *Journal of Development Effectiveness*, 1(3):217-226. <https://doi.org/10.1080/19439340903114636>

- Rutkowski, D. & Sparks, J. 2014. The new scalar politics of evaluation: An emerging governance role for evaluation. *Evaluation*, 20(4):492-508. <https://doi.org/10.1177/1356389014550561>
- Scriven, M. 1994. Evaluation as a discipline. *Studies in Educational Evaluation*, 20(1):147-166. [https://doi.org/10.1016/S0191-491X\(00\)80010-3](https://doi.org/10.1016/S0191-491X(00)80010-3)
- Stem, C.; Margoluis, R.; Salafsky, N. & Brown, M. 2005. Monitoring and evaluation in conservation: A review of trends and approaches. *Conservation Biology*, 19(2):295-309. <https://doi.org/10.1111/j.1523-1739.2005.00594.x>
- Taylor, D. 2005. Governing through evidence: participation and power in policy evaluation. *Journal of Social Policy*, 34(4):601-618. <https://doi.org/10.1017/S0047279405009177>
- White, H. 2010. A contribution to current debates in impact evaluation. *Evaluation*, 16(2):153-164. <https://doi.org/10.1177/1356389010361562>
- World Health Organization. 2015. *The health of the people: What works – the African Regional Health Report 2014*. Brazzaville: Regional Office for Africa, World Health Organization.