Speaking truth to power: Exploring a Ministry's evaluation department through evaluators' and policymakers' eyes

Levelt, L.; Pouw, N.

DOI
10.1177/13563890221109620

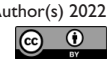
Publication date
2022

Document Version
Final published version

Published in
Evaluation : The International Journal of Theory, Research and Practice

Article

Evaluation
2022, Vol. 28(3) 379–395



DOI: 10.1177/13563890221109620
journals.sagepub.com/home/evi


Speaking truth to power: Exploring a Ministry's evaluation department through evaluators' and policymakers' eyes

Lotte Levelt
University of Amsterdam, The Netherlands

Nicky Pouw
University of Amsterdam, The Netherlands

Abstract

'Evidence-based' development policy has caused impact evaluations to prioritise accountability over addressing processual learning questions. Moreover, evaluation scholarship is dominated by surveys, whereas qualitative research remains scant. This article traces one particular evaluation, within the independent Evaluation Department of the Dutch Ministry of Foreign Affairs. It asks, 'How do evaluators and policymakers interact and what adjustments follow from the illustrative evaluation?' It used participant observations, documents and interviews with policymakers and evaluators. An in-depth thematic analysis resulted in a typology of evaluator roles: (1) knowledge broker, (2) facilitator, (3) archive, (4) truth-revealing and (5) critical voice. Finally, policymakers and managers adjusted in three ways: symbolic, instrumental and empowerment. These results imply that if evaluators deliberate a suitable role, they (1) increase their partial understandings of the programme under scrutiny and the involved stakeholders, and (2) enhance the potential of synergies in collective learning to emerge in an evaluation team and the broader institution.

Keywords

development policy, evaluation, institutional learning, policy adjustment, qualitative methods

Introduction

Background and problem statement

A major buzzword in current International Development practice and academia is 'evidence-basedness' (White and Raitzer, 2017). Following discussions of aid effectiveness of the 1990s

Corresponding author:
Lotte Levelt, Institute for Interdisciplinary Studies, Faculty of Science, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.
Email: l.levelt@uva.nl

and 2000s, a shared recognition has emerged among scientists and professionals that learning and accountability should be central concerns (Doucouliagos and Paldam, 2008; Easterly, 2007). Banerjee and Duflo (2011) famously pioneered these concerns in their experimental poverty research. One straightforward way in which actors and institutions in the International Development sector attempt to be (more) evidence-based is through evaluation of policies and programmes. However, the rise of 'evidence-based' Development Cooperation policy has caused evaluations to overemphasise accountability at the cost of learning (Kogen, 2018). This tension, between learning (i.e. reflecting on past programmes in hopes of improving these) and accountability (showing the ways in which taxpayer money is spent), is commonly referred to as the 'dual purpose' of evaluation. What is more, it is found that the goal of accountability often overshadows learning purposes of evaluations (Bjørkdahl et al., 2017).

One reason for this is that quantitative studies, such as randomised controlled trials (RCTs), can demonstrate direct impacts of programmes, while the benefits of qualitative research focused on policy learning are much less easily measurable and interpretable; these unfold over time and emerge from complex factors and stakeholders interacting in the process of programme implementation (Slade et al., 2020). As such, quantitative evaluations tend to focus on accountability between donors, implementing organisations and beneficiaries, overlooking the *learning* purpose that evaluations also intend to serve (Kogen, 2018). Rarely is the eventual uptake of lessons, drawn from evaluations, analysed. Furthermore, many studies in the policy and learning realm are survey-based. This means that current scholarship lacks detailed processual descriptions of learning processes between individuals. Moreover, many studies focus on cases where learning *did* happen, which skews our perception of policy adjustment (Moyson et al., 2017). This relates to so-called 'survivorship bias', where many studies in the evaluation realm analyse cases in which an evaluation led to a (desired) policy change, but instances where nothing happened, or an undesired change occurred, are rarely studied. Survivorship bias is widespread in, but not limited to, the world of business advice; stories of commercial success (either of individuals or businesses) are often distorted by ignoring all those who dropped out of college, or business ideas that never made it. Taleb (2010) refers to these unstudied cases of failure as 'silent evidence'. Similarly, one could make the argument that participants in evaluations are often times the 'usual suspects', creating further bias by leaving the 'unusual suspects' out of sight (see also Ware, 2014). In a recent special issue on policy success and policy failure, Dunlop (2017) calls attention to the importance of studying failure:

> Compared to the large volume of publications on 'good practices' and 'best practices', far less scholarly attention has been paid to 'bad practices' or 'worst practices' despite their widespread prevalence. As a result, public officials have failed to learn valuable lessons from these experiences. (p. 4)

As Dunlop states, analysing cases where learning did not happen (or policy failures) is important, not least because failures may prove a breeding ground for learning, according to May (1992):

> Cases involving policy failure are useful to consider since failure serves as a trigger for considering policy redesign and as a potential occasion for policy learning. One of the basic tenets of the organisational learning literature is that dissatisfaction with program performance serves as a stimulus for a search for alternative ways of doing business . . . Policy successes might be said to

> provide a stronger basis for learning by making it possible to trace conditions for success. However, dissatisfaction serves as a stronger stimulus for a search for new ideas than success. (p. 341)

In short, policy learning scholarship is dominated by survey-based research and its focus on policy success skews our perception of policy learning.

A recent study by Pattyn and Bouterse (2020) stresses the importance of focusing on interactions between policymakers and evaluations in learning processes. They find that engaging policymakers in the evaluation design increases evaluation use (Pattyn and Bouterse, 2020). Finally, Barbrook-Johnson et al. (2020) show that views of evaluators influence evaluation practice. For instance, the variety of backgrounds that evaluators come from lead to different conceptions of what constitutes an evaluation in the first place (Barbrook-Johnson et al., 2020). Hence, this study asks the question, 'How do evaluators and policymakers interact and what, if any, adjustments follow from the illustrative evaluation?'

This study focuses on learning (rather than accountability), using a mix of qualitative methods. It is focused on the position of evaluators and their interaction with policymakers. Finally, it analyses the adjustments made by policymakers and their managers, by following an illustrative evaluation as-it-happened. Because the study's data collection took place as the evaluation process unfolded, the subsequent policy changes were not yet known. In this way, the study avoided the tendency of focusing on usual suspects and stories of successful policy change. In short, this article aims to address the following knowledge gaps:

- Addressing the lack of processual qualitative studies in policy learning scholarship by researching the interactions between evaluators and policymakers, and
- Refocusing attention from accountability to institutional learning by analysing the follow-up of an unfolding evaluation process.

Theoretical framework

In order to situate this study within current policy evaluation scholarship, this section will first discuss institutional learning. Second, it provides an overview of existing evaluation uses, a metric used to analyse learning. Third and finally, it sheds a light on the positions of policymakers and evaluators.

Institutional learning

An important source for understanding policy change and learning is Hall's 1993 article 'Policy Paradigms, Social Learning and the State'. Hall distinguishes between three potential ways in which states change policies. A first-order change refers to changing levels of existing instruments (e.g. tax rates increase by *x%*). A second-order change involves the changing of instruments themselves (e.g. providing tax cuts instead of subsidies). A third-order change appears when the overarching goals, or paradigm, of policies change (e.g. moving from a Keynesian paradigm to monetarism). These third-order changes happen rarely and are often the result of political and societal contestation (Hall, 1993). This framework is useful because it sheds light on the spheres of influence and dimensions of change of policymakers and evaluators. They are visualised in the policymakers' sphere of influence in the conceptual scheme. For instance, evaluation departments often suggest rethinking of strategies behind policies,

Table 1. Types of evaluation use and learning found in policy evaluation scholarship.

Use	Definition	Learning	Source
Instrumental	Evaluation informs policymaking	Learning might take place, but mostly within pre-existing knowledge structures	Alkin and Taut (2003); Ledermann (2012)
Conceptual	Evaluation changes understanding of underlying assumptions, concepts, paradigms, etc.	Learning takes place, sometimes changes existing knowledge structures	Alkin and Taut (2003); Ledermann (2012)
Symbolic, tactical	Evaluation is used to legitimise or defend a previously held stance; used to postpone a decision	None	Alkin and Taut (2003); Ledermann (2012)
Accountability	Evaluation is used to determine how money was spent and whether goals were met	None	Azzam and Levine (2015)
Empowerment	Evaluation helps people to change their work, helps them address issues they're facing	Learning might take place	Fetterman (1994)

Source. This table was adapted from Bouterse (2016: 12). It was available for use for research purposes, as stipulated in the 'License to inclusion and publication of a Bachelor or Master thesis) in the Leiden University Student Repository' (Leiden University, 2012).

but the extent to which that is possible depends, in part, on the credibility and consistency of the status quo paradigm vis-à-vis an alternative one. Hall's conceptualisation of change and learning is used to analyse policy evaluation outcomes in this study.

Evaluation use

Government-commissioned evaluations are expected to not only serve accountability, but also stimulate institutional learning. As such, practitioners are 'utilization-focused', implying that evaluations are constructed with a specific user in mind and valued according to their usefulness (Patton, 2011: 315). *Evaluation use* is an often-used indicator for learning and Bouterse (2016) finds a total of five types of evaluation uses (see Table 1).

Especially instrumental, conceptual and empowerment use are relevant, for this is when learning takes place (Bouterse, 2016). In order to understand the variety of ways in which evaluations may be used, it is important to take a closer look at their users (policymakers) and creators (evaluators).

Policymakers and evaluators

It is advisable to analyse policymakers and evaluators at the individual level, since they are best positioned to describe their own changes in learning. In a recent study, Schmidt-Abbey et al. (2020: 205) call for an increased need to focus on evaluators themselves, given their 'embeddedness within an evaluand'. Grob (1992) studied policymakers and evaluators, which according to him sometimes appear to be worlds apart. He characterises evaluators as critical and concerned, and eager to make a difference, yet often ending up frustrated when their

findings are ignored or misused. Policymakers, on the contrary, complain that evaluations are too long, published too late or at times irrelevant (Grob, 1992).

Policymakers and evaluators therefore have separate spheres of influence (see Figure 1). Nonetheless, Pattyn and Bouterse (2020) show that their interaction may result in improved uptake of evaluation lessons. What is more, increased cooperation (e.g. developing a research question together, holding regular feedback interviews) between policymakers and evaluators may benefit learning through a process called developmental evaluation, or adaptive evaluation (Patton, 2011: 305). Hence, it is worthwhile to study the interaction between evaluators and policymakers, visualised in Figure 1.

Conceptual scheme: Key concepts and operational definitions

The conceptual scheme in Figure 1 guides the analysis of this study by highlighting its key concepts and relationships, showing an evaluation process. It will be used to structure the analysis of the study when presenting its results. Given the variety of contextual factors at play, it will be impossible to establish a causal relationship, hence the exploratory nature of this study. Nonetheless, a number of key concepts will be disentangled, and their relationships analysed. The main concepts of this study are evaluation, evaluandum (object of evaluation) and adjustment. On one hand, the study aims to analyse the position of the evaluator and their interactions with policymakers. This part of the study finds itself in the evaluators' sphere of influence. On the other hand, it analyses the interactive learning process of policymakers and evaluators by tracing the managerial adjustments following the illustrative evaluation.

Methodology

Research setting

Empirical data collection took place within the Dutch Ministry of Foreign Affair's Evaluation Department. This is a relevant research setting for three reasons: First, carrying out research here ensured access to rich qualitative data (e.g. Terms of Reference and interviews) which improved the robustness of the study. Second, the Evaluation Department is one of the first government evaluation units (founded in the 1970s) of development aid, resulting in a long tradition of evaluation expertise and high level of 'maturity' (Pattyn and Bouterse, 2020). As such, the Netherlands has a strong evaluation culture (Dahler-Larsen and Boodhoo, 2019). Third and finally, the setting provides the researcher with the opportunity of studying evaluation and policymaking 'as it occurs', increasing the ecological validity of the study. As the day-to-day business of policymaking is included in the analysis, the study paints a rich description of learning processes. This study's units of analysis include evaluations, evaluators and policymakers. The units of observation are employees of the Evaluation Department, policymakers of the Ministry of Foreign Affairs and evaluation reports.

Data collection

To answer the question of this research, the following data sources were used: three evaluation reports (ranging from development cooperation to foreign trade and international relations–themed studies), semi-structured interviews with evaluators and policymakers ($N=38$) and meeting minutes as well as participant observations in six stakeholder meetings, where

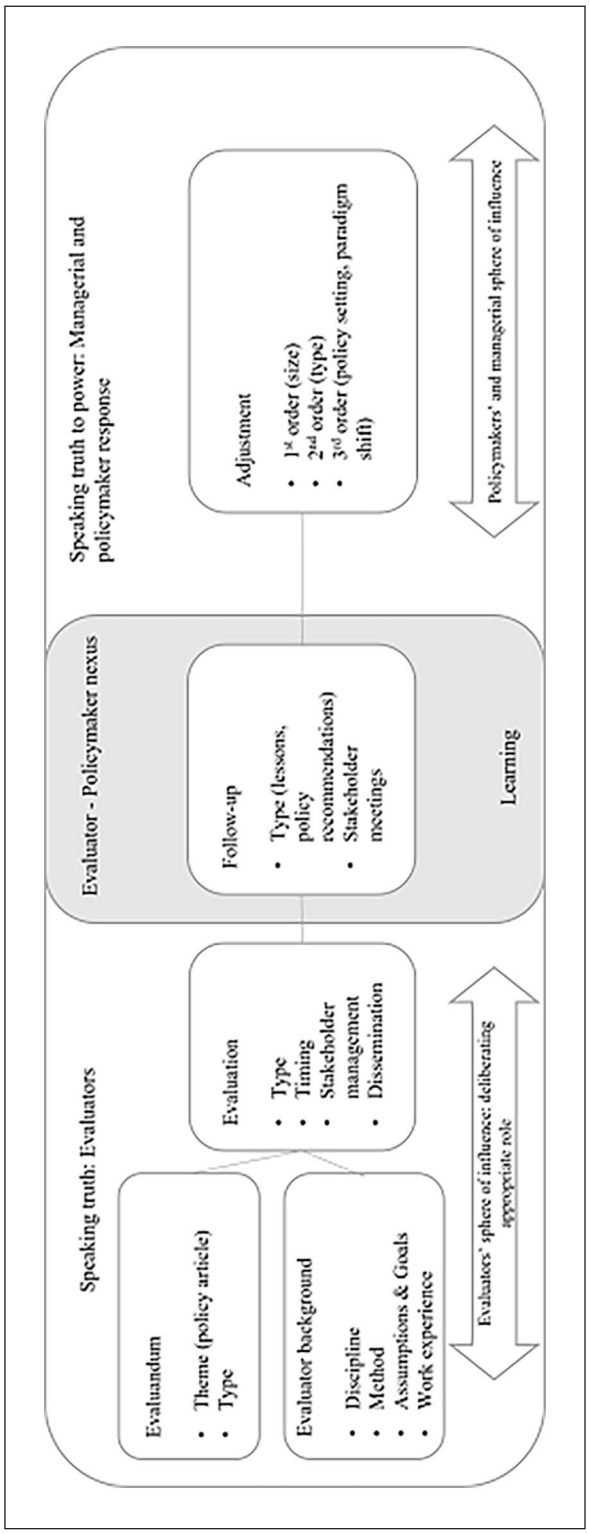


Figure 1. Conceptual scheme detailing the chronological process of an evaluation trajectory and policymakers and evaluators' spheres of influence.
Note. Authors' construction.

evaluation outcomes were discussed. Data collection took place in the period September–December 2019. For the semi-structured interviews, an interview guide was used to collect views and experiences of policymakers and evaluators, based on the evaluator sphere of influence of the conceptual scheme. Questions included 'In what discipline were you [evaluator] trained?' and 'What goal(s) do you [evaluator] try to achieve by carrying out/supervising evaluations?', while the interviews investigating the policymaker–evaluator nexus included, for instance, 'How do you [policymaker] estimate the influence of evaluation recommendations on policymaking generally?'. The goal of these observations and interviews was to move past 'official recordings' of actions, such as policy letters, and shed light on learning as experienced by individuals. Besides empirical data, this study makes use of existing literature and policy documents.

The illustrative evaluation process, used to study learning specifically, concerns the publishing and response to the report 'Less Pretension, More Realism' (Directie Internationaal Onderzoek en Beleidsevaluatie (2019a)). It is referred to as the 'illustrative evaluation' for the remainder of the article. Using a snowballing sampling technique, interviews were held with evaluators and policymakers, including the author of the policy response and the director of the respective policy department (Directie Internationaal Onderzoek en Beleidsevaluatie (2019b).

All interview transcripts, meeting minutes and documents were uploaded to Atlas.ti, coded using two cycles (starting with hypothesis coding, ending with evaluation coding) and subsequently thematically analysed. A detailed overview of the collected data can be found in Supplementary Table S1.

Limitations and data quality

This section shortly lists potential limitations and assesses its data quality. The study cannot infer causality, as there is no way of establishing a counterfactual, that is, what would have happened in a given situation if there had not been an evaluation. Moreover, it must be emphasised that the adjustments that follow evaluations are not per se *due* to the evaluation; an evaluation's input serves as one of many sources for policymaking and programme design.

To decrease selection bias among interviewees, all employees of the evaluation department were interviewed and posed the same questions to increase replicability in different thematic fields, or in other locations (Bryman, 2012; LeCompte and Goetz, 1982). Focusing on one evaluation department provides limited external validity (Bryman, 2012; LeCompte and Goetz, 1982). In this study, data collection took place in a mature evaluation setting. With decades of experience, this department has built a strong reputation and extensive knowledge of past, current and future programmes. Hence, the study's findings and recommendations may not be generalised to just any evaluation setting, but may prove relevant for other mature evaluation contexts.

Results

This section presents the main results of the analysis along two spheres of influence of the conceptual scheme. The model also portrays the illustrative evaluation process. First, it presents the position of evaluators and, second, it illustrates policymakers' adjustments in response to the illustrative evaluation. A full overview of the variety of data collected (interviews, participant observations and documents) for this study can be found in Supplementary Table S1.

Speaking truth? Evaluators play different roles and are uniquely positioned

In the semi-structured interviews with policymakers and evaluators, respondents were asked about their perceptions of evaluators. A number of themes recurred in the interviews surrounding questions about their perceived impact as well as their position within the Ministry.

A number of assumptions and views surrounding what evaluators ought to do, or not do, became apparent. For instance, several respondents indicated the evaluation department is too academic, as it desires to be 'the expert'. As one policymaker put it, 'The evaluation department has the tendency to want to come up with new methods, and first becoming experts in a domain rather than using existing material and moving ahead' (policymaker, interviewee 30, 2019). Interestingly, respondents held contrasting views about how critical evaluators should be. Several respondents indicated evaluators need to be more critical, as the evaluation department is precisely the department that can afford to do so, because its reputation and budget is strong. As such, it should not shy away from writing critical reports. It differs from consultancy and nongovernmental organisation (NGO)-based research: It suffers less from positive bias, which arises when evaluators over-report positive findings (or even exclude negative ones), in order to uphold a good relationship with the organisation funding the evaluation. Other respondents, on the contrary, urged evaluators to strike a more diplomatic tone: 'Evaluators need to avoid "attacking" policymakers by writing more diplomatically. Though there is a risk of writing too diplomatically; this requires pedagogic skills' (evaluator, interviewee 27, 2019).

Furthermore, several notions of the relationship between policymakers and evaluators surfaced from the interviews. A recurring concern among policymakers and evaluators alike was the apparent divide in understanding of each other's context:

> It is important for evaluators to understand the limits (in terms of workload, political sensitivity) of policymakers, and what their spheres of influence are. For instance, a recommendation to increase capacity is applauded by employees, but at the same time, they cannot decide to hire people themselves. (Policymaker, interviewee 30, 2019)

Besides their perceived lack of understanding, there is certainly a sense of appreciation for each other's work: Policymakers speak highly of evaluators and acknowledge their independent position:

> I also tell them (fellow policymakers, red.) to, when in doubt, ask IOB (the evaluation department) for advice, they can be seen as neutral experts, and their advice only sharpens conclusions we as policymakers draw about an evaluation. The reputation of IOB is high, both in the Netherlands and abroad. (Policymaker, interviewee 30, 2019)

Finally, the expert status is recognised by policymakers, who indicate there is a recent desire to improve monitoring and evaluation (M&E) capacity in several departments: 'At the same time, I think now, there's more desire for having ex-evaluators in policy departments, because evaluators have time, unlike policymakers, to get really deeply informed with a topic, which means they become almost experts' (policymaker, interviewee 30, 2019).

In summary, respondents hold a variety of views regarding the position of evaluators. On the basis of the interview data presented above, it was found that various, and at times

Table 2. A typology of evaluation department's roles, based on interviews with evaluators and policymakers ($N = 38$) in 2019 at the Ministry of Foreign Affairs.

Evaluator as . . .	Characteristics of this role	Types of products	Discussion of pros and cons
Knowledge broker	Motivating, are knowledgeable of state-of-the-art evidence in their theme	Systematic reviews, evidence gap maps	Dependent on seniority and credibility of the evaluator Independence is ensured as only information is handed (not choices, arguments or advice)
Advisor, facilitator	Encourages policymakers to request advice, organises stakeholder sessions	Workshops, lectures and one-on-one meetings, recommendations in evaluation reports	On the boundary of independence Straightforward way to stimulate learning on an individual level through personal interaction
Ministerial archive, memory	Actively looks back at what has been done (both within the evaluation department and within Ministry); to avoid reinventing wheel; taking records	Policy reviews, syntheses	One of the only departments that looks back, which is valuable especially given constant shuffling of staff in Ministry
Truth-revealing	Investigative, seeks Parliament connection	Short articles which may be 'spicy', or evaluations about topical themes (e.g. Turkey deal evaluation)	Creates attention
Critical voice	Focused on accountability, aims to improve status quo	Critical evaluations, lessons, perhaps even sanctions	Evaluators have an independent status, can afford to be critical. However, may scare off stakeholders.

contrasting, functions were attributed to the evaluation department. To this end, a typology was created of roles, characteristics, outcomes and a discussion of their advantages and disadvantages. This typology is presented in Table 2.

These various roles are within the sphere of influence of the evaluator and may therefore serve as a deliberation tool. If evaluators are conscious of their respective roles, within the team and institution, they become more aware of their acquired understandings and the partiality and potential complementarity of that. The specific implications of the typology will be discussed in the 'Implications and recommendations for M&E practitioners' section.

Finally, the interview data comprised many views of the interactions between policymakers and evaluators. This policymaker–evaluator nexus, where varying types of evaluation use surfaced, and hence learning may take place, will be discussed in the next section.

Speaking truth to power? Policymakers and managers adjust in various ways

This section presents the results of interviews conducted with policymakers and evaluators, as well as a document analysis (i.e. the evaluation report and policy response letter), all pertaining to one illustrative evaluation trajectory. Three different types of evaluation use (symbolic, instrumental and empowerment) were found and will be discussed below.

Symbolic. Evaluators found that the achievement and sustainability of results had been impaired by high levels of fragmentation: Funding was spent too scarcely between various small and geographically distant activities. The policy response letter of the Cabinet (signed by the Minister of Development Cooperation and Trade) recognised this recommendation. The Ministry asserted it has started limiting the number of activities, as more focus will increase the quality of Dutch efforts in development cooperation (Directie Internationaal Onderzoek en Beleidsevaluatie (2019b)).

During the interviews, several policymakers indicated that this lesson is not new: Fragmentation had been a recurring issue in development cooperation spending. However, two policymakers did point out that the document for policymakers to 'make their case' better for reducing fragmentation, vis-à-vis their managers, but also towards implementing organisations like NGOs. As such, the evaluation is used as a substantiation for the ongoing fragmentation discussion within the Ministry.

Instrumental. In response to recommendations, a number of tangible actions have been taken: first, the establishment of an internal working group for defragmentation efforts as well as exploring alternatives to tendering, which include important so-called 'change agents' within the Ministry. The goal of this group is to investigate the existing bottlenecks in defragmentation efforts and to find the best way to reduce the number of activities of departments by about 30 per cent. It was one of the first instances that a dedicated working group was established after an evaluation, thus setting up the stage for a 'learning-team' in which collective learning could come to full fruition.

Empowerment. Evaluators find an overemphasis on accountability vis-à-vis learning in current M&E efforts. The use of standardised indicators is justified, but its dominance damages the use of M&E for learning purposes. Policymakers face pressure from Parliament to report results. As a consequence, result frameworks developed in advance hardly suit the changing and fragile contexts in which programmes take place. In this way, both NGOs and the Ministry are not incentivised to reflect and learn, or report negative results, either fearing the loss of funding or facing parliamentary criticisms (Directie Internationaal Onderzoek en Beleidsevaluatie (2019a)).

The Cabinet acknowledges that monitoring and evaluation should be given more attention across the board. Hence, it promises to increase the capacity for M&E staff as well as training current employees, both within the Ministry and at embassies (Directie Internationaal Onderzoek en Beleidsevaluatie (2019b)).

In interviews, policymakers recognise the tentative rising interest in M&E across the Ministry. There appears to be more room to do something around 'lessons learnt' and M&E. One policymaker thought that, on one hand, external pressures, like politicians asking for transparency about results, drive this development. On the other hand, she observed an internal drive to organise M&E better, although this differs per subject and level: 'At the activity-level, there is a lot of opportunity for change and amendment. It gets trickier at higher levels, where political wishes may run counter lessons we learn about effectiveness' (policymaker, interviewee 31, 2019).

Summarising, the results of the illustrative evaluation trajectory showed a variety of adjustments and interactions between policymakers and evaluations. Symbolic (evaluation is used as substantiation in internal discussions about fragmentation), instrumental (the goal of 30%

activity reduction and establishment of a working group) and empowerment (call to increase staff capacity in the ministry) uses of evaluations were found. These findings are presented in Table 3, adapted from Bouterse (2016), which recaps evaluation uses and presents an illustration from this evaluation trajectory.

Implications and recommendations for M&E practitioners

This penultimate section takes the study's key findings and, based on their implications, formulates a number of recommendations to M&E practitioners. A snapshot of these findings, implications and recommendations can be found in Table 4.

As reported in the 'Results' section, two key findings were distilled from the study's data.

First, evaluators play different roles and are uniquely positioned. The typology of roles, presented in Table 2, gives an idea of these roles, typical characteristics and corresponding products. This is not the first study to challenge the idea of evaluators as singularly oriented to research methods and models. Skolits et al. (2009) find that evaluators take on a wide variety of demands and recommend a more 'situational' perspective on the role of the evaluator. They find that consideration of the expected evaluation activities, their particular demands and required products (e.g. types of deliverables) warrants careful consideration of roles when recruiting evaluation team members (Skolits et al., 2009). Therefore, this study recommends deliberation of required roles at the very outset of an evaluation trajectory. However, role deliberation is by no means definitive. Evaluators may, where possible, take on multiple roles throughout an evaluation trajectory. Verwoerd et al. (2020) find that combining the role of evaluator and facilitator, for instance, resulted in an evaluation that better matched the project under scrutiny. This flexibility in roles can provide an evaluation with emergent qualities, where adjustments can be made in response to needs of policymakers, (external) researchers or changing political realities (Verwoerd et al., 2020). Hence, the benefit of an evaluation trajectory with emergent qualities, that allows evaluators to change roles when necessary. Furthermore, the study found that evaluators are deemed independent and having time to get deeply involved in a project. Grob (2012) shows that while decisions in policymaking are never made by one person or organisational entity, evaluators have a unique position because of their independent and helpful reputation. What is more, the nature of their work allows evaluators to build their knowledge, since they have the time to get deeply acquainted with programmes under scrutiny, as well as state-of-the-art research of 'what works' (Tourmen et al., 2021). Their unique, independent position, as well as the time they have to build a strong basis of knowledge, implies their added value lies with acting as knowledge brokers while recognising the partiality of their own knowledge and need for knowledge exchange with others.

Second, three types of managerial adjustments were found when analysing the illustrative evaluation trajectory: symbolic, instrumental and empowerment. A more detailed overview of these adjustments was presented in Table 3. Managers may use evaluations in a symbolic way, for instance, to substantiate an already ongoing discussion. This entails a risk of evaluators being pressured to report previously held beliefs (Pleger and Sager, 2018). However, Pleger and Sager find that these influences are not necessarily negative, but may also be positive. They offer three differentiating questions to evaluators to discern the type of influence at hand: Is the attempt to influence consciously or unknowingly (*awareness*)? Is the reason of influence self-interest or an attempt to improve the quality of the evaluation (*intention*)? And finally, is the influence in accordance with scientific standards (*accordance*)? (Pleger and Sager, 2018).

Table 3. Results: Types of evaluation use and illustrations of learning found in case study data (evaluation reports ($N = 5$), participant observations ($N = 17$) and interviews ($N = 38$) with the Ministry of Foreign Affairs in 2019).

Use	Definition	Learning	Present in case study?	Illustration
Instrumental	Evaluation informs policymaking	Learning might take place, but mostly within pre-existing knowledge structures	+/–	• Target to reduce no. of activities by about 30% • Investigation into alternatives to tender procedures ongoing • Establishment of working group • Shifting of funds from the ministry to embassies, to increase suitability to local context.
Conceptual	Evaluation changes understanding of concepts, paradigms	Learning takes place, sometimes changes existing knowledge structures	–	
Symbolic, tactical	Evaluation is used to defend a previously held stance; used to postpone a decision	None	+	• Evaluation is used as substantiation for ongoing fragmentation discussion within the Ministry • The evaluation also caused the postponement of the 'maatschappelijk middenveld' debate (Ministry wanted to await evaluation findings and lessons before publishing new subsidy framework)
Accountability	Evaluation is used to determine how money was spent and whether goals were met	None	+	• Evaluation showed the effects and results of the two evaluated programmes
Empowerment	Evaluation helps people to change their work, helps them address issues they're facing	Learning might take place	+	• Recommendations to increase M&E capacity (no. 3) as well as cut the number of activities (no. 1) is applauded by policymakers • The general lack of staff as an organisational issue is a recurring theme

Note. Authors' construction, adapted from Bouterse (2016: 12).

Table 4. Key findings, implications and resultant recommendations.

Finding	Implication	Recommendation
Main finding I: Evaluators play different roles and are uniquely positioned		
Typology of roles	Range of activities and demands lead to different roles to be assumed (Skolits et al., 2009)	Reflect on necessary roles that should be fulfilled in an evaluation team; how can these be complementary?
	When evaluators take on multiple roles, for example, as facilitators and evaluators, this enhanced understanding of the evaluated programme and involved stakeholders (Verwoerd et al., 2020)	Incorporate emergent qualities (where different roles can be assumed throughout the trajectory) and potential for knowledge synergies
Neutral, independent and reputable status	Evaluators' independent reputation makes them valuable and stand out from other professionals involved in policymaking (Grob, 2012)	Added value of having time and credibility to act as knowledge broker while being conscious of the boundaries of knowledge linked to evaluators' respective roles
Time to get deeply involved in a topic	Evaluators theorise and build experience because they have time to do so (Tourmen et al., 2021)	
Main finding II: Policymakers and managers adjust in various ways		
Symbolic (evaluation was used as substantiation in ongoing discussions)	Pressure to report previously held beliefs and external influence can be negative and positive (Pleger and Sager, 2018).	Know how to discern different influences to effectively manage these
Instrumental (working group and 30% reduction in activities)	Instrumental use by managers increases as recommendations are included (Bourgeois and Whynot, 2018)	Write down actionable recommendations and validate these with stakeholders
Empowerment (more attention and capacity for M&E)	Empowerment evaluation is a promising tool for (evaluation) capacity building (Donaldson, 2017)	Maintain sensitivity to policymaker context by including stakeholders proactively

Hence, evaluators need to know how to distinguish positive and negative external influences to manage these effectively. Furthermore, Bourgeois and Whynot (2018) assert that instrumental use of evaluations, by managers specifically, increases as actionable recommendations are included. Therefore, this study recommends evaluators to do just that. Finally, existing studies corroborate this study's finding that the empowerment use of an evaluation may be promising; Donaldson (2017) argues that empowerment evaluation has always prioritised stakeholder involvement, as well as stimulate evaluation capacity (not only of evaluators, but with policy departments and implementing organisations), leading to increased use of evaluations (Donaldson, 2017). Hence, this study recommends evaluators to proactively manage stakeholders (e.g. by involving them in the trajectory from the outset) and maintaining (in)formal contact with policymakers to increase and maintain sensitivity to their context.

Discussion

This section highlights key contributions of the study and subsequently outlines potential avenues for future research.

The first key result, that evaluators play different roles, was summarised in Table 2. The idea of *evaluator roles* is not new; indeed, Skolits et al. (2009) previously defined evaluator roles on the basis of their demands. However, an empirical basis for this conceptualisation was lacking to date. This study has provided this empirical basis, by distilling roles from a mix of empirical data sources, ranging from semi-structured interviews to participant observations. Of these roles, that of knowledge broker is expected to be most effective, since evaluators' added value lies with their time to get familiar with programmes, as well as their independent reputation (Grob, 2012; Tourmen et al., 2021). Ridde (2007: 1020) sees the evaluator as knowledge broker as '. . . an intermediary between the worlds of research and action'. The interviewed evaluators of this study indeed previously worked in academia or in policy departments and implementing organisations, such as NGOs. This mix of backgrounds, and unique position between research and action, requires careful consideration of roles required in particular evaluation teams. This study finds that a mix of evaluator roles, as well as incorporating emergent qualities in evaluation trajectories, where roles may switch, increases understanding of the programme under scrutiny.

Second, three types of managerial adjustments were found in response to the illustrative evaluation trajectory. Certainly, the Ministry has, in response to this evaluation, taken concrete actions, summarised in Table 3. Examples of these include the postponement of a parliamentary debate (the Ministry wanted to await evaluation findings and lessons before publishing the new subsidy framework), the target to reduce activities by about 30 per cent and the goal of increasing M&E capacity and cutting the number of activities per policymaker. Simultaneously, these illustrations highlight three types of evaluation use (see Table 1), corresponding with Bouterse's (2016) overview of evaluation uses: symbolic, instrumental and empowerment use of evaluations. Interestingly, the illustrative evaluation process shows resemblances with Hall's (1993) fundamental framework for policy change. For instance, the goal of reducing fragmentation, to which policy departments have responded by initiating 30 per cent cuts in activities, portrays a first-order change, a mere decrease in the level or 'setting' of an instrument. Furthermore, the suggestion to move away from tendering as method of contracting implementing organisations portrays a second-order change, the changing of instruments. Finally, the typical recommendation for evidence-based programmes hints at a paradigmatic change. This also portrays the lively debate surrounding 'what works', in which evaluators have a role to play as knowledge brokers, is alive and well. As Hall (1993) points out, in the realm of first- and second-order change, there is room for expert judgement. The paradigm, however, provides the context in which potential adjustments are made. These are not directly amenable because they refer to reigning worldviews and are the result of political contestations, determining, for instance, who is deemed an expert. Although evaluators can hardly influence the dominant paradigm, a government can look at evaluation departments for inspirations and input about alternative, perhaps better, paradigms than the status quo. In this combination, of looking back and reflecting, but also offering alternative ways of thinking and acting, lies the worth of an evaluation department.

In conclusion, we believe this study's empirically based typology of evaluator roles constitutes a novel contribution to policy learning scholarship. These roles call for careful consideration of evaluation teams and incorporation of emergent qualities in evaluation trajectories. The role of knowledge broker is promising, since evaluators' time and reputable status gives them credibility and extensive insight into programmes. Managers adjust to evaluations in various ways. Yet, evaluators are equipped to respond to potential pressures

by knowing how to discern positive from negative influences, as well as by engaging proactively with stakeholders. Finally, the study contributes to an ongoing methodological gap in evaluation literature identified by Moyson et al. (2017). Using a mix of qualitative methods, and analysing an evaluation as-it-happened, the study presents unprecedented insights into evaluation processes within a Ministry.

Several suggestions for future research arise from this article. A replication study could be executed in another context, for instance, in the Ministry of Foreign Affairs of another country, which may have different organisational structures, or within another Dutch Ministry. It would be interesting to analyse whether follow-up and learning work through similar mechanisms in other policy areas. Future studies could incorporate elements of systems thinking and institutional analyses to discern bottlenecks and path dependencies in policy learning. Furthermore, in terms of methodology, future research could use time series methods, to analyse whether evaluations' recommendations stick in the long-term, or comparative studies to analyse the follow-up of several evaluations, instead of one illustrative evaluation. Finally, future studies could dig deeper into the enabling circumstances for learning, in order to move closer to the ideal of 'evidence-based' policymaking. An example research question could be, 'What factors incentivise, or constrain, policymakers to learn from evaluation?' There's a lot to learn.

Acknowledgements

The authors thank Caspar Lobbrecht, Rob van Poelje and Wendy Asbeek-Brusse of the Evaluation Department (Ministry of Foreign Affairs of the Netherlands) for their guidance throughout the fieldwork period. The authors thank Hebe Verrest for her commentary in early stages of the study and Sarah Gane for her French translation.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The study was independently financed by the authors.

ORCID iD

Lotte Levelt https://orcid.org/0000-0001-5569-2604

Supplemental material

Supplemental material for this article is available online.

References

Alkin MC and Taut SM (2003) Unbundling evaluation use. *Studies in Educational Evaluation* 29(1): 1–12.

Azzam T and Levine B (2015) Politics in evaluation: Politically responsive evaluation in high stakes environments. *Evaluation and Program Planning* 53: 44–56.

Banerjee A and Duflo E (2011) *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: PublicAffairs, 7–9.

Barbrook-Johnson P, Proctor A, Giorgi S, et al. (2020) How do policy evaluators understand complexity? *Evaluation* 26(3): 315–32.

Bjørkdahl K, McNeill D and Reinersten H (2017) Confronting the contradiction – an exploration into the dual purpose of accountability and learning in aid evaluation. Report for the Swedish Expert Group for Aid Studies (EBA). Available at: https://eba.se/wp-content/uploads/2017/05/Webbversion-FINAL-blanksida.pdf (accessed 21 March 2022).

Bourgeois I and Whynot J (2018) The influence of evaluation recommendations on instrumental and conceptual uses: A preliminary analysis. *Evaluation and Program Planning* 68: 13–8.

Bouterse M (2016) *Explaining evaluation use: A qualitative comparative analysis of factors influencing instrumental use of evaluations*. MSc Thesis, Leiden University, 12. Available at: https://studenttheses.universiteitleiden.nl/access/item%3A2629771/view (accessed 21 March 2022).

Bryman A (2012) *Social Research Methods*, 4th edn. Oxford: Oxford University Press, 45–8.

Dahler-Larsen P and Boodhoo A (2019) Evaluation culture and good governance: Is there a link? *Evaluation* 25(3): 277–93.

Directie Internationaal Onderzoek en Beleidsevaluatie (IOB) (2019a) *Less pretension, more realism: An evaluation of the reconstruction programme, the strategic partnerships in chronic crises programme and the addressing root causes tender process*. Report for the Ministry of Foreign Affairs, 1 July. The Hague: Ministry of Foreign Affairs, 9–14.

Directie Internationaal Onderzoek en Beleidsevaluatie (IOB) (2019b) *Beleidsreactie – Evaluatie Wederopbouw- En Spcc-programma's En Arc-tenderproces*. Report for the Ministry of Foreign Affairs, 2 October. The Hague: Ministry of Foreign Affairs, 2–4.

Donaldson SI (2017) Empowerment evaluation: An approach that has literally altered the landscape of evaluation. *Evaluation and Program Planning* 63: 136–7.

Doucouliagos H and Paldam M (2008) Aid effectiveness on growth: A meta study. *European Journal of Political Economy* 24(1): 1–24.

Dunlop CA (2017) Policy learning and policy failure: Definitions, dimensions and intersections. *Policy & Politics* 45(1): 3–18.

Easterly W (2007) *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. New York: Oxford University Press, 43–7.

Fetterman DM (1994) Empowerment evaluation. *Evaluation Practice* 15(1): 1–15.

Grob GF (1992) How policy is made and how evaluators can affect it. *Evaluation Practice* 13(3): 175–83.

Grob GF (2012) Evaluators in a world of valuators. In: Julnes G (ed.) *Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation. New Directions for Evaluation*, vol.133. Chichester: Wiley, 91–6.

Hall PA (1993) Policy paradigms, social learning, and the state: The case of economic policymaking in Britain. *Comparative Politics* 25(3): 275–96.

Kogen L (2018) What have we learned here? Questioning accountability in aid policy and practice. *Evaluation* 24(1): 98–112.

LeCompte MD and Goetz JP (1982) Problems of reliability and validity in ethnographic research. *Review of Educational Research* 52(1): 31–60.

Ledermann S (2012) Exploring the necessary conditions for evaluation use in program change. *American Journal of Evaluation* 33(2): 159–78.

Leiden University (2012) License to inclusion and publication of a Bachelor or Master Thesis in the Leiden University Student Repository. Available at: https://studenttheses.universiteitleiden.nl/handle/1887/license%3A1 (accessed 21 December 2021).

May PJ (1992) Policy learning and failure. *Journal of Public Policy* 12(4): 331–54.

Moyson S, Scholten P and Weible CM (2017) Policy learning and policy change: Theorizing their relations from different perspectives. *Policy and Society* 36(2): 161–77.

Patton MQ (2011) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: Guilford Press, 305–15.
Pattyn V and Bouterse M (2020) Explaining use and non-use of policy evaluations in a mature evaluation setting. *Humanities and Social Sciences Communications* 7(85): 19.
Pleger L and Sager F (2018) Betterment, undermining, support and distortion: A heuristic model for the analysis of pressure on evaluators. *Evaluation and Program Planning* 69: 166–72.
Ridde V (2007) Are program evaluators judges and/or knowledge brokers? *Journal of Epidemiology and Community Health* 61: 1020.
Schmidt-Abbey B, Reynolds M and Ison R (2020) Towards systemic evaluation in turbulent times – Second-order practice shift. *Evaluation* 26(2): 205–26.
Skolits GJ, Morrow JA and Burr EM (2009) Reconceptualizing evaluator roles. *American Journal of Evaluation* 30(3): 275–95.
Slade R, Hazell P, Place F, et al. (2020) Evaluating the impact of policy research: Evidence from the evaluation of rural policy research in developing countries. *Evaluation* 26(4): 541–61.
Taleb N (2010) *The Black Swan: The Impact of the Highly Improbable*, 2nd edn. New York: Random house, 101.
Tourmen C, Berriet-Solliec M and Lépicier D (2021) The spontaneous theoreticians: How evaluators build and revise their knowledge of programs through experience. *Evaluation* 27(3): 307–25.
Verwoerd L, Klaassen P, Van Veen SC, et al. (2020) Combining the roles of evaluator and facilitator: Assessing societal impacts of transdisciplinary research while building capacities to improve its quality. *Environmental Science & Policy* 103: 32–40.
Ware A (2014) Beyond the usual suspects: Complexity of fragility and analytical framework. In: Ware A (ed.) *Development in Difficult Sociopolitical Contexts*. London: Palgrave Macmillan, 3–23.
White H and Raitzer D (2017) *Impact Evaluation of Development Interventions: A Practical Guide*. Manila, Philippines: Asian Development Bank, 1–3.

Lotte Levelt is at Institute for Interdisciplinary Studies, Faculty of Science, University of Amsterdam, the Netherlands. Levelt's research interests include research and policymaking in international development, rethinking economics, interdisciplinarity, sustainability and degrowth, global inequalities, circular economy and mixed-method research.

Nicky Pouw is at Amsterdam Institute for Social Science Research, University of Amsterdam, the Netherlands. Pouw's research interests include economics of wellbeing in the context of inclusive development, addressing issues of economic redistribution, inequality, poverty, marginalisation, sustainability and voice and empoerment.