Developing evaluation standards and assessing evaluation quality

Authors:
Mike Leslie[1]
Nishendra Moodley[1]
Ian Goldman[2]
Christel Jacob[2]
Donna Podems[3]
Mark Everett[2]
Terence Beney[3]

Affiliations:
[1]Palmer Development Group (PDG), South Africa

[2]Department of Planning, Monitoring and Evaluation, South Africa

[3]Independent Consultants, South Africa

Correspondence to:
Mike Leslie

Email:
mike@pdg.co.za

Postal address:
PO Box 46830, Glosderry 7702, South Africa

Dates:
Received: 09 Feb. 2015
Accepted: 08 June 2015
Published: 20 Aug. 2015

How to cite this article:
Leslie, M., Moodley, N., Goldman, I., Jacob, C., Podems, D., Everett, M. *et al*., 2015, 'Developing evaluation standards and assessing evaluation quality', *African Evaluation Journal* 3(1), Art. #112, 13 pages. http://dx.doi.org/10.4102/aej.v3i1.112

The article explains the rationale for the development of standards for evaluation practice, the process followed in developing those standards, and how those standards inform the quality assessment of evaluations. Quality assessment of evaluations are conducted as a routine activity of the South African National Evaluation System (NES). The importance of quality assessment for improving the state of evaluation practice in South Africa is illustrated by presenting results from the quality assessments undertaken to date. The paper concludes by discussing the progress on the development of a public Evaluations Repository to manage and provide access to completed evaluations and their quality assessment results, and offering some concluding analytical remarks.

Introduction

The Department of Planning, Monitoring and Evaluation (DPME) in the South African Presidency is the custodian of the national monitoring and evaluation (M&E) system. Prior to 2011, work on evaluation in South Africa was sporadic. There was no established national evaluation system, no common approaches and no set standards were applied. In November 2011 a National Evaluation Policy Framework (NEPF) was approved by Cabinet and DPME began implementing the National Evaluation System (NES). Part of this work has been to facilitate national priority evaluations as stipulated in the National Evaluation Plan (NEP). The NEP is approved by Cabinet, and all reports are submitted to Cabinet. This is discussed further in the article on the NES.

Other work involved setting up the elements of the NES that apply not just to NEP evaluations but to evaluations across government, such as standards and competences for evaluations. One of these elements has been a quality assessment system for government evaluations, informed by the standards and competences developed by government. All NEP evaluations completed to date have been subjected to quality assessment.

Background

Considering international perspectives on evaluation standards and quality

In 2012, DPME commissioned a research paper that explored the development of evaluation standards to support the implementation of the NES. The paper aimed to generate a basis for discussion and shared understanding of evaluation standards, and the related evaluation competences that are needed to effectively undertake, commission and use evaluation within the South African government (King & Podems 2014).

The qualitative exploratory research relied on a desk review of published evaluation standards and in-depth interviews with South African government officials, civil society members and academics. In addition, English-speaking evaluators in other countries (Australia, New Zealand, Canada, Great Britain, and the United States) who had experience with developing evaluations standards, guidelines or evaluator competences were consulted. Informed by this research, and further guided by the principles and values stated in the NEPF, a draft evaluation standards document was produced in June 2012 (Podems 2012).

The eventual draft Evaluation Standards drew heavily from several guidelines, including the Joint Committee on Standards for Educational Evaluation (JCSEE), Program Evaluation Standards (JCSEE 1994) referenced by both the Canadian Evaluation Society (CES 2012) and the American Evaluation Association (AEA), the African Evaluation Association (AfrEA) Standards, standards developed by the German *Gesellschaft für Evaluation e.V*. (DeGEval), and the Swiss Evaluation Society (SEVAL) (Podems & Podems 2014).

The JCSEE, a coalition of major professional associations concerned with the quality of evaluation, developed a set of standards for the evaluation of educational programmes based on the concepts

of utility, feasibility, propriety and accuracy. The AEA recognised programme evaluation standards developed and refined by the JCSEE in 1981, 1994 and 2011. In 2011 the JCSEE added a group of standards on accountability. Interviews suggest that this was a tumultuous process and that there were many detractors to the additional standards (Podems 2012).

The AfrEA Evaluation Guidelines (AfrEA 2006) were developed in 2006 and built on the JCSEE standards. They include 35 standards that are divided into four major principles:

- *Utility*, for produced information and expected and provided results.
- *Feasibility*, for realism, cautiousness and efficiency.
- *Respect of ethics*, respect of legal and ethical rules.
- *Precision and quality*, for a relevant methodology related to the goal and the subject matter of the evaluation.

The 25 DeGEval standards (DeGEval 2002) are arranged across the four original JCSEE categories, as are the Swiss standards. The Swiss adaptation of the JCSEE standards generalises their application from education to a diversity of content areas, an adaptation also employed by the DeGEval standards. The Swiss standards were the only set reviewed that specifically mention three areas that DPME was interested in targeting, namely the evaluators themselves, those who commission the evaluations, and other persons participating in the evaluation (Windmer, Landert & Bachmann 2000). The SEVAL standards also provide guidance on how their standards should be used, acknowledging that it is not always possible to meet each standard equally, and that they need to be adapted to the particular context. However, decisions on the application of standards need to be transparent and clearly documented. This qualified flexibility implies that a certain level of evaluation knowledge and skill is required to appropriately inform decisions related to the application of standards.

Another interesting aspect of the SEVAL standards is the use of a 'Functional Overview' that maps out what evaluation activity requires which evaluation standards. Functional areas include the decision to conduct an evaluation, defining and planning the evaluation, collecting and analysing the information, evaluation reporting and budgeting, concluding an evaluation contract, managing the evaluation, and personnel and evaluation. This approach is echoed in the structure of the standards endorsed in February 2010 by the Development Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development (OECD). The aim of these standards '… to improve quality and ultimately to strengthen the contribution of evaluation to improving development outcomes' (DAC 2010:5), appears particularly relevant to the South African case. The structure of the OECD-DAC standards includes the categories of overarching considerations, namely, purpose, planning and design, implementation and reporting, follow-up, and use and learning.

One standard not discussed or mentioned in the literature reviewed was a standard that addressed equity, which is of interest given the emphasis on gender-responsive and equity focused evaluations by international groups such as EvalPartners. In addition, whilst there is substantial discussion on the need for evaluation in Africa to be more contextually embedded (Tarsilla 2014), the only apparent example is that of the AfrEA standards.

The standards review demonstrated that there were two main models to draw from, that is, the quality criteria based approach of the JCSEE and its derivatives, or the phase based approach used by the DAC, and suggested in the functional overview of SEVAL. The JCSEE standards are based on years of specialised and focused discussions and have been referenced and applied for more than a decade and the AfrEA standards were adapted from this model (pre 2011). Alternatively, the SEVAL standards suggested a useful approach to present standards by grouping them into functional categories. At the same time they recognise that standards need to be adapted to each situation, and that not all standards have the same weight in any given situation. In addition, the DAC standards offer a different model that also appears relevant given the functional phased approach and cross-cutting considerations in South Africa, and touching on many core issues.

The South African standards

Based on a review of these different approaches, it was decided in July 2012 at a workshop with DPME and the South African Monitoring and Evaluation Association (SAMEA) that the most useful framework for South Africa was the DAC approach because of its functional application across phases, and provision for overarching considerations that could be interwoven throughout an evaluation. South African standards were drafted and first published in August 2012.

Later the same year, DPME facilitated several participatory public forums to gather feedback and refine the standards document. The first forum presented the standards to key stakeholders, mainly from government and academia, with some representation from civil society. This initial gathering achieved consensus on the structure and content of the standards document. DPME then circulated the document through various forums to gain wider stakeholder feedback. In this process, the DPME partnered with SAMEA to facilitate three national and provincial workshops. DPME then used this feedback to refine the standards document and published this document on their website for general use (DPME 2014). The NES has drawn on the Evaluation Standards to inform the national evaluation guidelines, government focused training material, and suggested government evaluation templates and tools.

Summarising the standards

Overarching considerations

In the final draft of the South African standards (DPME 2014), the DPME added an introductory section that addressed

overarching evaluation considerations such as a partnership approach, evaluation ethics, quality control mechanisms, and others. Largely drawn from the DAC standards (2010), this section raises key issues that need to be considered throughout the evaluation process. The following are the seven thematic areas taken as a direct excerpt from the document:

- *Partnership approach:* In order to increase ownership of the evaluation and maximise the likelihood of use, and build mutual accountability for results, a partnership approach to development evaluation is considered systematically early in the process.
- *Free and open evaluation process*: Where appropriate the evaluation process [*should be*] transparent and independent from programme management and policy-making, to enhance credibility.
- *Evaluation ethics*: Evaluations abide by relevant professional and ethical guidelines and codes of conduct for individual evaluators [*and are*] undertaken with integrity and honesty … respect[*ing*] human rights and differences in culture, customs, religious beliefs and practices of all stakeholders.
- *Coordination and alignment*: To … improve co-ordination of evaluation and implementation of evaluation results, the evaluation process must take into account the roles of different stakeholders, seeking to ensure those critical to the intervention are involved in the evaluation ...
- *Capacity development*: The process of evaluation [*should have a*] positive effect on the evaluation capacity of the partners involved as well as developing the capacity of evaluators. … This capacity development should be through an explicit learning-by-doing process, as well as in the process adopted.
- *Quality control*: Quality control [*should be*] exercised throughout the evaluation process. … quality control is carried out through an internal and/or external process. Peer review… [*and*] an evaluation quality assessment (EQA) [*should*] be conducted to reflect on the process as well as the product of the evaluation …
  - *Project management*: The evaluation [*should be*] conceived, planned, managed and conducted and results… made available to commissioners in a timely manner to achieve the objectives of the evaluation … Changes in conditions, circumstances, timeframe and budget [*should be*] reported … [*and*] explained, discussed and agreed between the relevant parties. (DPME 2014)

Evaluation phases and standards

DPME then elected to break the quality standards down into four phases, with the overarching considerations blended across them. The four phases and the comprising standards are listed below:

- **Phase 1 - Planning, Design and Inception** incorporates the following standards: clarity of purpose and scope in the terms of reference (ToR); evaluability of the programme and adequate resourcing for the evaluation. This phase also considers stakeholder involvement, governance and management structures, the selection of evaluation service provider as well as the inception phase of the evaluation.
- **Phase 2 - Implementation considers the following standards:** The independence of the evaluator where necessary, key stakeholder involvement, relevant ethical considerations, and the implementation of the evaluation within allotted time and budget.
- **Phase 3 - Reporting addresses the following standards:** Dealing with intermediate reports; writing and presentation; report formatting considerations; coverage of the report including evaluation questions answered; context of the development intervention; the intervention logic; explanation of the methodology used; clarity of analysis of conclusions; acknowledgement of changes and limitations of the evaluation; validity and reliability of information sources, as well as acknowledgement of disagreements within the evaluation team. Finally this phase also considers the incorporation of stakeholders' comments.
- **Phase 4 - Follow-up, use and learning:** The final phase of evaluations outlined in the standards for evaluation in government includes the following standards: Timeliness, relevance and use of the evaluation, the systematic response to and follow-up on recommendations, dissemination of evaluation results, as well as reflection on the evaluation process and product (DPME 2014).

Within each of these four phases is a set of evaluation standard items, totalling 74 standards across all phases. The overarching considerations, phases and standards therefore served as the basis for developing a quality assessment tool. This tool, along with an accompanying framework, was then developed into an online quality assessment system.

Quality assessment system

Assessment tools and scoring system

The assessment tool developed for the quality assessment of government evaluations, that is, the Evaluation Quality Assessment Tool (EQAT), has followed the structure and evolving content of the *Standards for Evaluation in Government* document as it was released in draft form and then in an adopted form in 2014 (DPME 2014). The tool is conceived to assess the entire evaluation process, inclusive of the three main role-players, in combination. In each of the phases mentioned above, a group of standards comprise sub-assessment areas within a phase. Phases receive scores based on a composite measure of a group of unique evaluation standards specific to that phase. The aggregates of all evaluation standard items in a phase are individually weighted to give a composite measure of the phase.

Each of the standards is scored using a Likert-type rating scale. These standards are rated on an interval scale ranging from very poor (1), inadequate (2), adequate (3), good (4) to excellent (5). Following the first two rounds of quality assessments, the need was identified to enhance the reliability of quality assessors' application of the scale, and a set of standard level definitions for each of the five levels is in development.

In rare instances when an evaluation standard does not apply for a given evaluation, a not applicable (N/A) rating is provided. However, the N/A is not a rating in the true sense, as it designates that the evaluation standard is omitted entirely from the composite measure of the criteria and phase.

In the case of the seven overarching considerations listed earlier, these cross-cutting assessment principles are applied over the four phases, and reflected in standards within the phase that are aligned directly to the overarching consideration. Thus, standards aggregated within a phase produce a phase score, but some of the same standards also align across phases to produce a score for each of the overarching considerations. In this way, a group of evaluation standard items from across all four phases can be combined differently to provide a measure of an overarching consideration.

When calculating an overall quality rating for the evaluation, it is recognised that the different phases are of different degrees of significance to the overall evaluation relative to the others. Thus, in producing an overall composite measure of all the evaluation standard items within each of the four phases (a quality assessment score), each of the phases is given a differential weighting based on the significance.

Table 1 illustrates the weighting applied to each phase for the two rounds of quality assessments. Round one was characterised by evaluations that occurred prior to the development of government standards and competences, when there was a significant lapse in time since completion of the evaluation and the evaluation report served as the primary evidence for assessment. Round two weightings reflect the importance of the phases as they currently stand and assume quality assessment within a short period of finalisation incorporating available documentary evidence and interviews with all key respondents. Thus, in calculating quality assessment scores between the two rounds, round one places disproportionate weight on the standards included under the Implementation and Report phases. In round two, the weighting of the phases shifts to a more even spread, whilst still concentrating at the largest proportional weighting on standards addressing the report phase.

Audit of government evaluations and sample

An audit exercise of existing government evaluations by the European Union-funded Programme to Support Pro-poor Policy Development (PSPPD) initially identified 135 evaluations conducted between 2005 and 2011. As DPME constituted its evaluation panel in 2012, a further 34 possible evaluations were identified which had been undertaken by panel members. However, on closer scrutiny of some of the reports, many appeared to be classic surveys, general research, compliance and performance audits, rather than having a distinct evaluative approach. In the end, using evaluation report availability as a condition, along with a set of criteria for determining the evaluative nature of the report, a set of 83 evaluations were included for quality assessment (DPME 2013a).

TABLE 1: Weighting of evaluation phases for historical evaluations.

Phase of evaluation	Weighting round 1	Weighting round 2
1. Planning and design	10	20
2. Implementation	30	20
3. Report	50	40
4. Follow-up, use and learning	10	20

In the second round, the anticipated set of 70 evaluations for quality assessment (40 national and 30 provincial evaluations) have not been forthcoming. As a result, the sample included for analysis includes only those 25 evaluations that have been quality-assessed to date. Of those 25 evaluations included for this analysis, 5 are NEP evaluations, 14 are national evaluations conducted outside of the NEP, and 6 are provincial evaluations. These are sampled on the basis of availability for the quality assessment at this time.

Quality assessment methodology

Document review

Assessors review all available evaluation documentation at the outset of a quality assessment. The minimum following documentation is generally sought and reviewed:

- Terms of Reference (ToR).
- Inception Report.
- Data Collection Tools or Instruments.
- Evaluation Report.

In addition, any other supporting documentation relevant to the evaluation process is also considered if available, including, but not limited to the proposal, meeting minutes, progress and draft reports, presentations, and so forth. ToRs and evaluation reports are consistently available, whilst the availability of inception reports and data collection instruments is variable for non-NEP evaluations.

Documents are considered as evidence relevant to the different phases of the evaluation (e.g. ToR to the planning and design phase, evaluation report to the reporting phase, etc.) and serve as an evidence base, in combination with qualitative data obtained by interview.

Stakeholder interviews

Assessors attempt to engage a minimum of three role-players for each evaluation. As per the *Evaluation Competency Framework for Government* (DPME 2012) the role players sought as respondents include, but are not limited to, the M&E advisor for the department or commissioning organisation, the programme manager or most relevant manager for the evaluand, and the evaluators. In the case of NEP evaluations, the DPME evaluation director is also as stakeholder required for interview.

Stakeholder interviews are conducted using a semi-structured interview guideline that reflects the phases,

criteria and standards applied in the EQAT, with an emphasis on obtaining information for those phases and overarching considerations not covered by the document review, such as the follow-up, use and learning. Questions are differentiated between the three types of respondents so as to triangulate perspectives and deepen the information available to inform the rating.

Assessment and reporting

Assessors (e.g. evaluation consultants, academics and government staff with significant evaluation experience) are responsible for using the evaluation documents available and the qualitative data obtained from the stakeholder interviews to synthesise and judge each government evaluation against 74 standard items on the 5 point scale (this increased from 67 standards in the first round and is currently under review). Assessors are responsible for providing supporting commentary to justify the score based on the available evidence for every standard.

An online web-based platform was developed by the consulting team as part of the first round to facilitate the quality assessment scoring, commenting, capturing, analysis and document management process (DPME 2013a).

Once all standards are completed and composite measures have been generated for each phase and overarching consideration, these scores form the analytical basis for writing quality assessment summaries that pronounce on the overall quality of the evaluation (DPME 2013a).

All quality assessments are moderated prior to finalising reporting. Moderation entails a review of the consistency, completeness and rigour of the quality assessment against all the 74 standard items based on the evaluation details, motivating commentary, scoring, overall summary and overall documentation and respondents. This seeks to ensure that the approach of different assessors is generally consistent and ensure inter-assessor reliability (DPME 2013a).

A quality assessment report is then generated to share with the key evaluation stakeholders (e.g. the DPME evaluation project manager, the participating department, and the evaluation team). A window of three weeks is provided for stakeholder comment. If comment is received then the quality assessment goes back to the assessor for revision in light of the feedback and any evidence received. Once final revisions are made or the three week window passes, quality assessments are considered final (DPME 2013a).

The overall quality assessment summary, supported by the ratings and commentary for each standard, together with the categorisation information about the evaluation and references for all source documentation and interviews, comprises the reporting content for each of the 25 government evaluations quality-assessed in this round.

Limitations of methodology

The quality assessment methodology has some limitations, not least that the tool itself adopts a 'one-size' approach to applying standards to the six different types of evaluations identified in the NEPF. This provides comparable measures for evaluations of varying degrees of sophistication (e.g. quasi-experimental impact and formative design or implementation evaluations).

Subjecting new completed evaluations to quality assessment has also rendered some standards developed in the first round inappropriate, notably those that assumed significant time had elapsed to demonstrate use. The time between completion of the evaluation and subjecting it to quality assessment has been significantly shortened, rendering it too soon to meaningfully pass judgement on the evaluation in terms of the follow-up, use and learning phase. However, this has prompted investigation of a possible later follow-up, use and learning assessment.

Inter-assessor reliability of scoring has been largely managed through the moderation period. However, arising from an ongoing need to provide greater guidance for those standards, five standard level definitions are being developed.

Lastly, the sample of evaluations is relatively small and uneven across government and is therefore not necessarily representative of all evaluations conducted in the period. Extrapolating the results must be treated with caution.

Results

Evaluations deemed to be of an adequate standard or above

Applying the minimum rating of 3 (deemed to be on average of an adequate evaluation standard) as the cut-off point for considering evaluations as an acceptable quality, 13 evaluations were assessed as falling below this standard of quality.

Figure 1 presents a ranked spread of total scores across the 25 evaluations sampled, whilst the colouration of the stacked bars indicates the contribution to the overall score by phase. Each phase contributed a proportion to the overall rating of an evaluation based on the round 2 weightings of the phase shown in Table 1.

Although the nature and type of evaluations varied considerably, evaluations rated well overall. The majority of those under quality assessment in round 2 exceeded the minimum threshold of 3 (Table 2).

In the first round the relatively highly score of 3.57 represented the average quality assessment rating. However, in round 2 this score declined slightly to 3.50 on average. The distribution of the total scores achieved by the rounds 1 and 2 evaluations is presented in the table below.

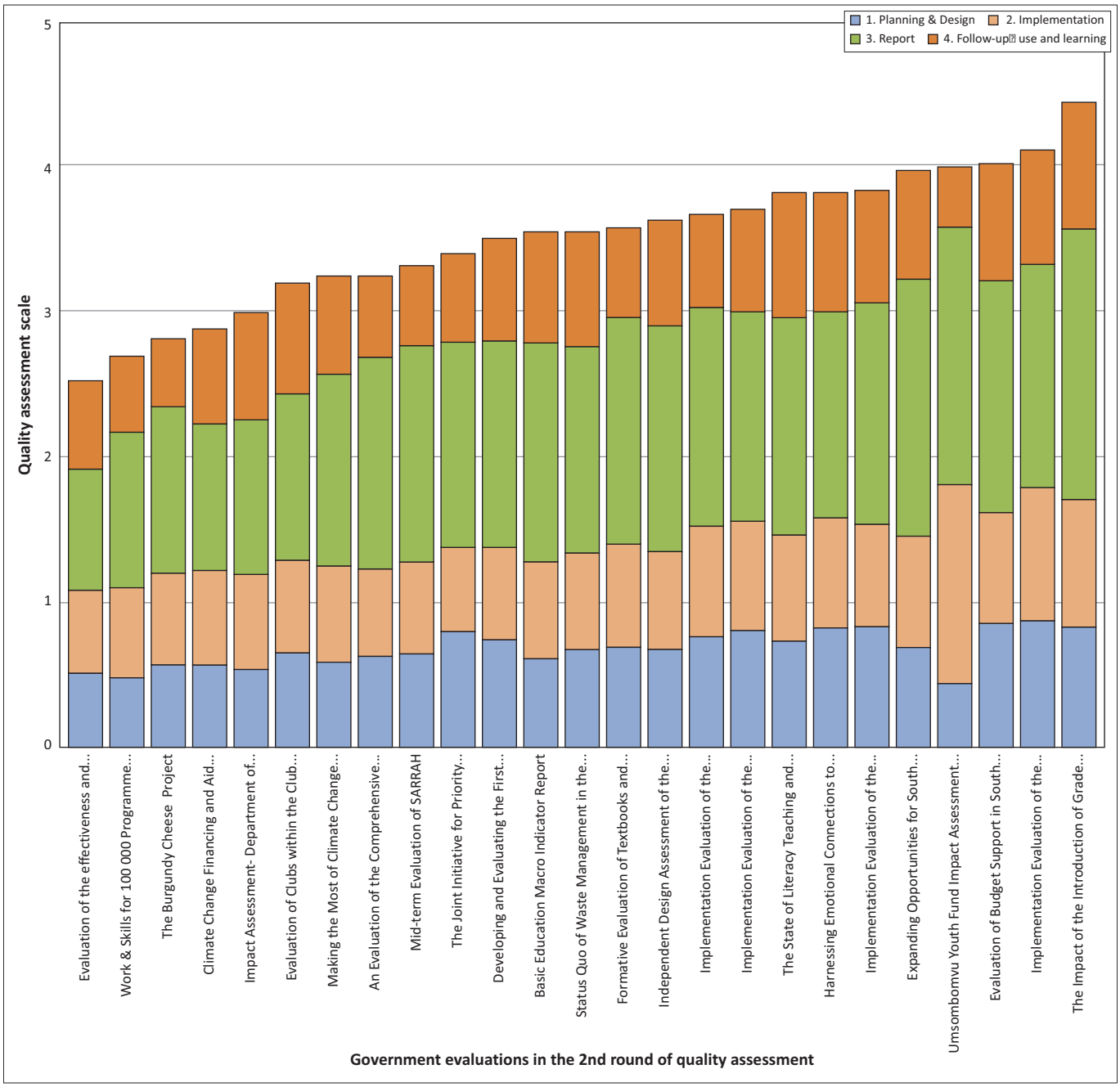

FIGURE 1: Ranked total scores of each evaluation with contributing components.

TABLE 2: Distribution of total scores in round 1 and round 2.

Score range†	Number of evaluations round 1	% of total	Number of evaluations round 2	% of total
< 2.0	0	0	0	0
< 2.5	1	1	0	0
< 3	12	14	5	20
< 3.5	14	17	5	20
< 4	40	48	11	54
< 4.5	15	18	4	16
≤ 5	1	1	0	0
Total	**83**	**100**	**25**	**100**

†, The score range in the table reflects the upper limit of each category, for ease of reading. The lower limit is not stated but is above the previous score range. For example the score range < 3.5 is meant to imply all evaluations that scored < 3.5 but who scored ≥ 3, the preceding upper limit.

The distribution of the data across the two rounds above indicates skewness above the mid-point of 3 on the rating scale in both. The spread of the data for round 2 indicates 10 evaluations (40%) of the sample scored below 3.5, whilst only 4 evaluations (16%) stand out as exemplifying a particularly high quality (Table 2).

Trends in evaluation type

Table 3 below presents the six types of evaluations classified in the NEPF: diagnostic, design, implementation, impact, economic and evaluation synthesis. From the following spread of data it is clear that the majority of government evaluations conducted within the round 2 sample fall under the more established evaluation types of implementation and impact evaluations. Less common in this regard were diagnostic, design and economic

TABLE 3: Distribution of evaluations by type.

Type	Number round 1	Number round 2
Diagnostic	8	4
Design	1	1
Implementation	38	16
Impact	37	7
Economic	6	1
Evaluation synthesis	6	0
Other	3	1
Total	**99**	**30**

evaluations whilst not a single evaluation synthesis was included in this sample.

Although implementation evaluations still predominate in round 2, there is a shift in the distribution away from evaluations considered impact assessments. Although the sample is limited, it reflects greater emphasis on formative assessments in round 2.

The totals in both rounds are higher than the number of evaluations subjected to assessment in each round. In this case 5 of the 25 evaluations were listed as multiple types, and so the total in Table 3 is 30. Some of these evaluations were completed almost before the supporting guidelines for the typology were available and these indications of type are retrospective approximations of best fit within the current policy framework.

The graph presented in Figure 2 illustrates the average aggregate rating across each of the evaluation typologies for the two rounds combined. This must be treated with caution as there were very few of some types of evaluation. Notable from this graph is that the lowest average score is held by implementation evaluations, which have the largest portion of the sample overall. Also interesting to note is that economic evaluations fared the best in terms of average score overall ($n = 7$).

Standard ratings for the planning and design phase

The graph below illustrates the spread of ratings for all evaluations with respect to each of the evaluation standards within the first phase: planning and design. In the case of 'not applicable' ratings, the bar chart is coloured white, creating the appearance of no distributions. However, as could be expected of the planning and design phase given the relatively small proportion of recent evaluations, there is a significant proportion of 'not applicable' ratings included from the first round.

The colour coding of the five point scale demonstrates those standards that fare poorly in terms of the frequency with which they scored 1s and 2s especially. In this regard, standard *1.4.1: There was explicit reference to the intervention logic or the theory of change of the evaluand and in the planning of the evaluation*, stands out as one of the least applied quality standards. This is particularly problematic because of the

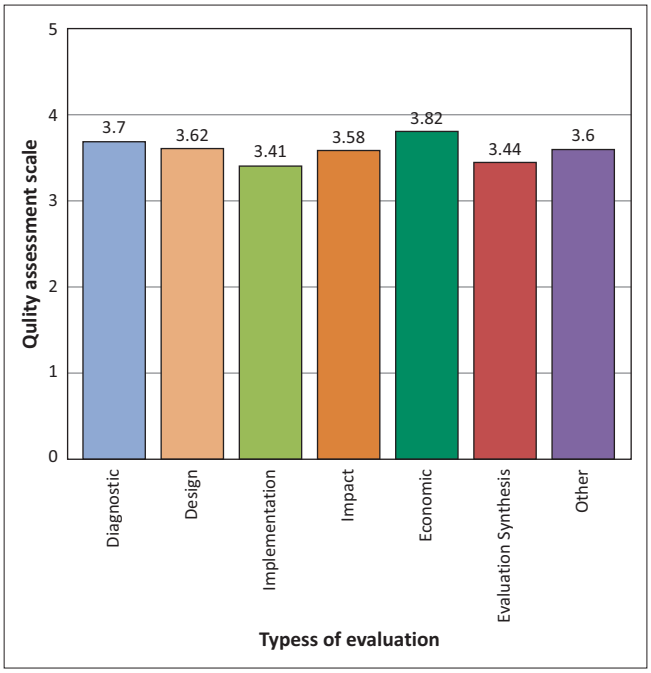

FIGURE 2: Average score of evaluation types.

requisite need to understand the programme theory of a policy, programme or project to credibly assess it.

Similarly, standards *1.2.4: Where appropriate, the evaluation planned to incorporate an element of capacity building of partners/staff responsible for the evaluand* and *1.4.5: There was a planned process for using the findings of the evaluation*, also received a significant portion of low ratings. In the case of the first standard, the common failure to appropriately plan for the use of the evaluation findings resonates later during the fourth phase of evaluations. Standard 1.4.5 reflects later during the overarching consideration discussion and will be addressed further there.

Figure 3 not only indicates shortcomings, but also highlights some areas of good practice in the historical evaluations work included as part of this sample. For instance, standard *1.4.3: The planned methodology was appropriate to the questions being asked*, consistently scored well, suggesting that the methodological intentions were appropriate for the kind of answers sought from the evaluation. Not surprisingly for the public sector, there was also evidence that the evaluations rated well in terms of standards *1.3.1: There was evidence that a review of the relevant policy and programme environments had been conducted and used in planning the research*, and to a lesser extent *1.3.2: There was evidence of a review of appropriate literature having been conducted and used in planning the research.*

Standard ratings for the implementation phase

The graph below presents the ratings for the standards covering the implementation phase. Of note are the three new standards introduced for the implementation phase in round 2. They account for the comparatively smaller stacks under standards *2.2.3: Where appropriate, the evaluation*

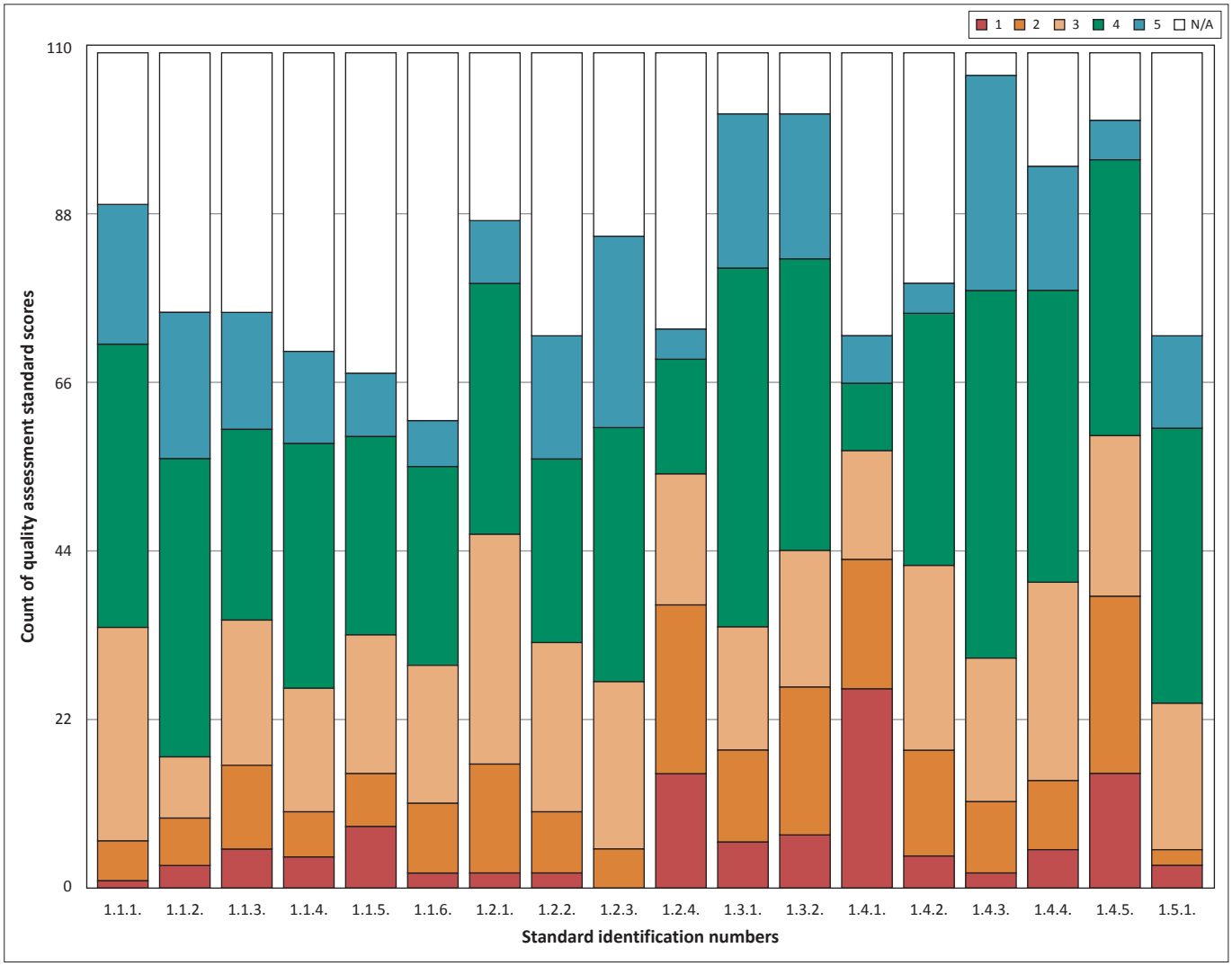

FIGURE 3: Distribution of ratings per standard: Planning and Design Phase.

team incorporated an element of skills development amongst the evaluators, 2.2.4: Peer review of the agreed evaluation design and methodology occurred prior to undertaking data collection, and *2.3.2: A pilot of data collection instrumentation occurred prior to undertaking data collection*. When compared with the previous graph, it is noticeable for the fewer number of standards within this phase, and the lesser proportion of 'not applicable' ratings.

Within this phase, standard *2.2.2: Where appropriate, an element of capacity building of partners responsible for the evaluand was incorporated into the evaluation process*, is clearly the biggest challenge. When considered with the three new standards, it is clear that preparatory review and capacity building stand out as shortcomings. Based on the emerging data, capacity building is something that has been under-provided for in the course of the evaluations despite it being policy for NEP evaluations (DPME 2011).

Many of the evaluations are also examples of good practice, as is the case for standards *2.1.3: The evaluation team was impartial and there was no evidence of conflict of interest* and *2.3.4: Forms of data gathering were appropriate given the scope of evaluation*, where the graph above illustrates the large proportion of evaluations receiving ratings of 4 and 5 for these two standards (Figure 4).

Standard ratings for the reporting phase

The graph below presents the ratings for all of the standards within the reporting phase. In this phase there are also three new standards as indicated by the three comparative short stacks. Comparatively fewer standards were 'not applicable' in this phase, as a result of the availability of the evaluation report as evidence is a prerequisite for undertaking the quality assessment. The standards could therefore be said to be biased in terms of historical application to the primary document on which the quality assessment was based.

There is one case within this phase where an evaluation standard was not applied by the quality assessor nearly half of the time, namely, for standard *3.4.4: Conclusions were drawn with explicit reference to the intervention logic or theory of change*. The result is that one of the critical standards for determining the overall credibility of the evaluation report

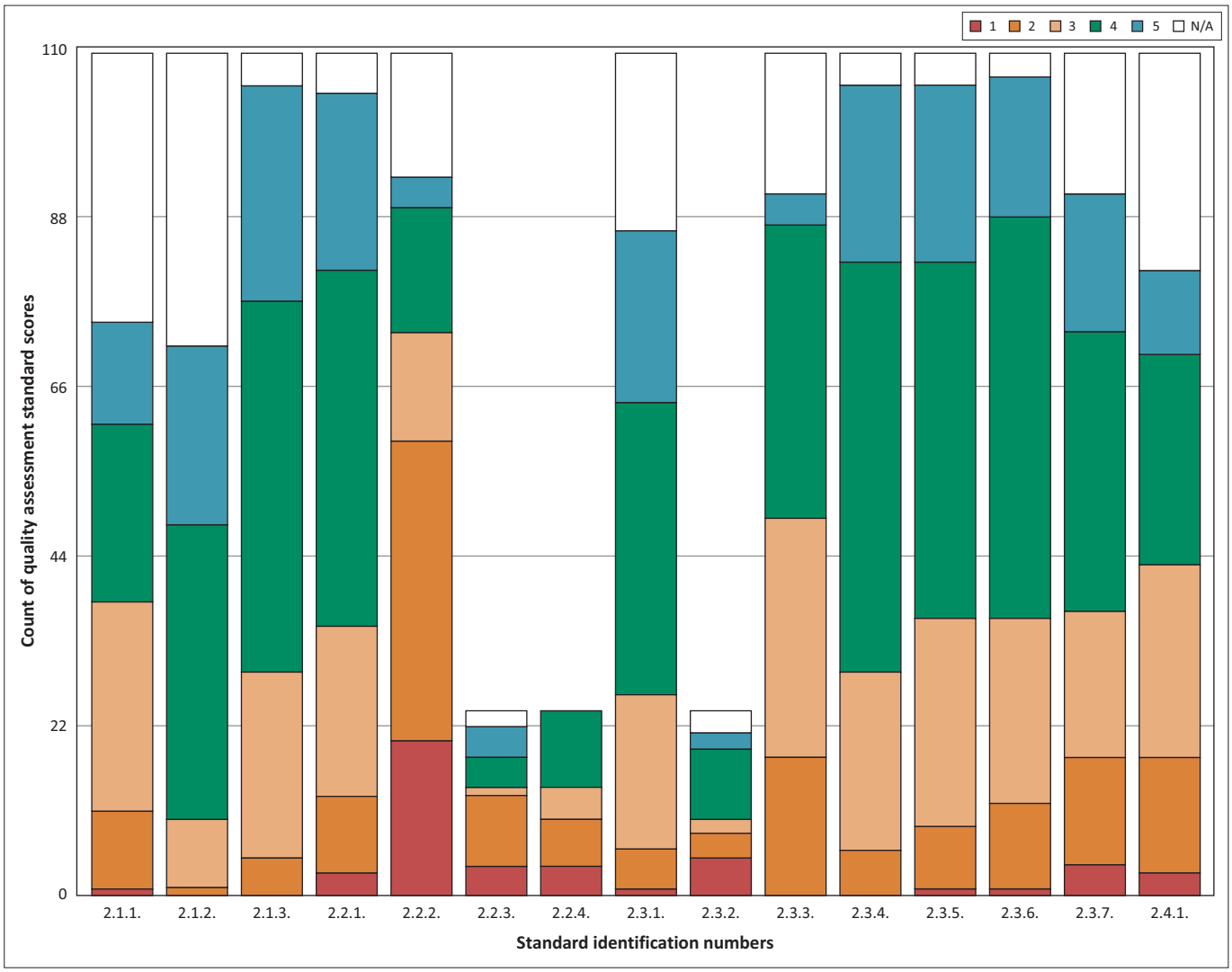

FIGURE 4: Distribution of ratings per standard: Implementation Phase.

in relation to the intervention under assessment is largely without evidence. This is an important finding and a matter that needs to be addressed more broadly within government evaluative work. In the meantime, all new NEP evaluations now require a theory of change to be developed as part of the evaluation process (DPME 2013b).

Standard *3.1.1: Executive summary captures key components of the report appropriately* stands out with *3.1.6: Acknowledgement of limitations of all aspects of the methodology and findings are clearly and succinctly articulated*, for the frequency with which the lowest rating is given. Similarly, *3.3.4: There was appropriate recognition of the possibility of alternative interpretations* and *3.3.6: Relevant limitations of the evaluation are noted*, received regularly low ratings of 1 and 2, indicating a consistent neglect of alternative interpretations of data and findings, as well as an acknowledgement of overall limitations. This practice is detrimental to the use of the evaluations, specifically because it has the potential of presenting one set of conclusions and recommendations to decision-makers only, when in fact there may be multiple and competing interpretations or options available. Low adherence to these standards does not suggest the analytical rigour and reflection needed for good, useful and robust evaluations (Figure 5).

In terms of an evaluation standard that rated well across all the evaluations, standard *3.6.3: There were no risks to participants in disseminating the original report on a public website*, exemplifies good practice. However, many of the reports included in the round 1 sample are already available online and may have been rated highly on account of being publicly available. Their inclusion in the sample is therefore directly biased by their availability, thus favouring them in terms of the application of the standards. This is hardly the case in round 2 and one would expect that as more recently completed evaluations are included in the sample for quality assessments that this standard would be reflective of a more representative sample of evaluations.

Standard ratings for the follow-up, use and learning phase

The graph below presents ratings for each standard in the fourth evaluation phase of follow-up, use and learning. Because these standards are rated based on information

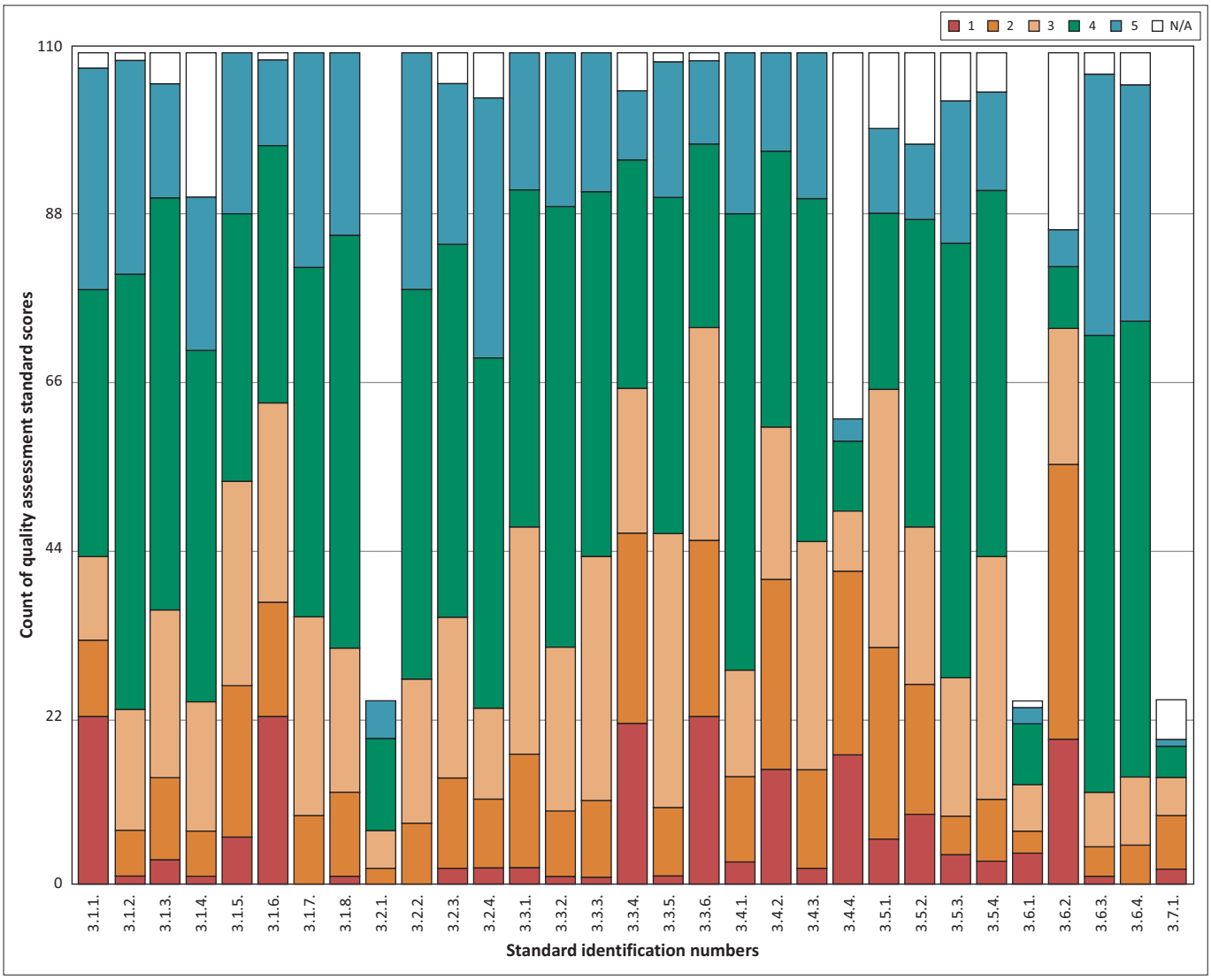

FIGURE 5: Distribution of ratings per standard: Implementation Phase.

obtained mostly after completion of the evaluation report, they rely heavily on interviews with evaluation role players and are therefore subjective and perception based. Further, because round 2 occurred in close proximity to the completion of the evaluation report, there has not always been sufficient time to allow for utilisation of the evaluation results.

It is clear that the standard *4.2.2: A reflective process has been undertaken by staff responsible for the evaluand to reflect on what could be done to strengthen future evaluations*, receives the lowest ratings within this phase. Standard *4.2.5: Development of a draft improvement plan has been started, but not completed, based on the findings and recommendations set out in the evaluation*, is the only new standard within this phase and has a proportionally high number of 'not applicable' ratings, suggesting that it may be applied too early or that the expectation that a separate planning document always succeeds an evaluation may not be appropriate.

Other standards that had some low ratings within this phase included *4.2.7: There was clear evidence of instrumental use - that the recommendations of the evaluation were implemented to a significant extent*, and *4.2.8: There was clear evidence that the evaluation has had a positive influence on the evaluand, its stakeholders and beneficiaries over the medium to long term*. Both of these standards have a significant portion of 'not applicable' ratings, especially in round 2, largely because of the time frames within which the evaluations are subjected to quality assessment.

The standard that is consistently rated highly within this phase is *4.2.6: The report is publicly available (website or otherwise published document), except where there are legitimate security concerns*. The high ratings are to be expected as the availability of the evaluation report was the first prerequisite for including the evaluation in the sample in round 1, and those that were readily available online were therefore more likely to be included in the sample because they could be easily accessed, whereas many other evaluation reports that were originally proposed for the sample were not included on account of the inability to access copies of the reports (Figure 6).

The round 2 quality assessments have highlighted some of the challenges of applying standards for this fourth phase within

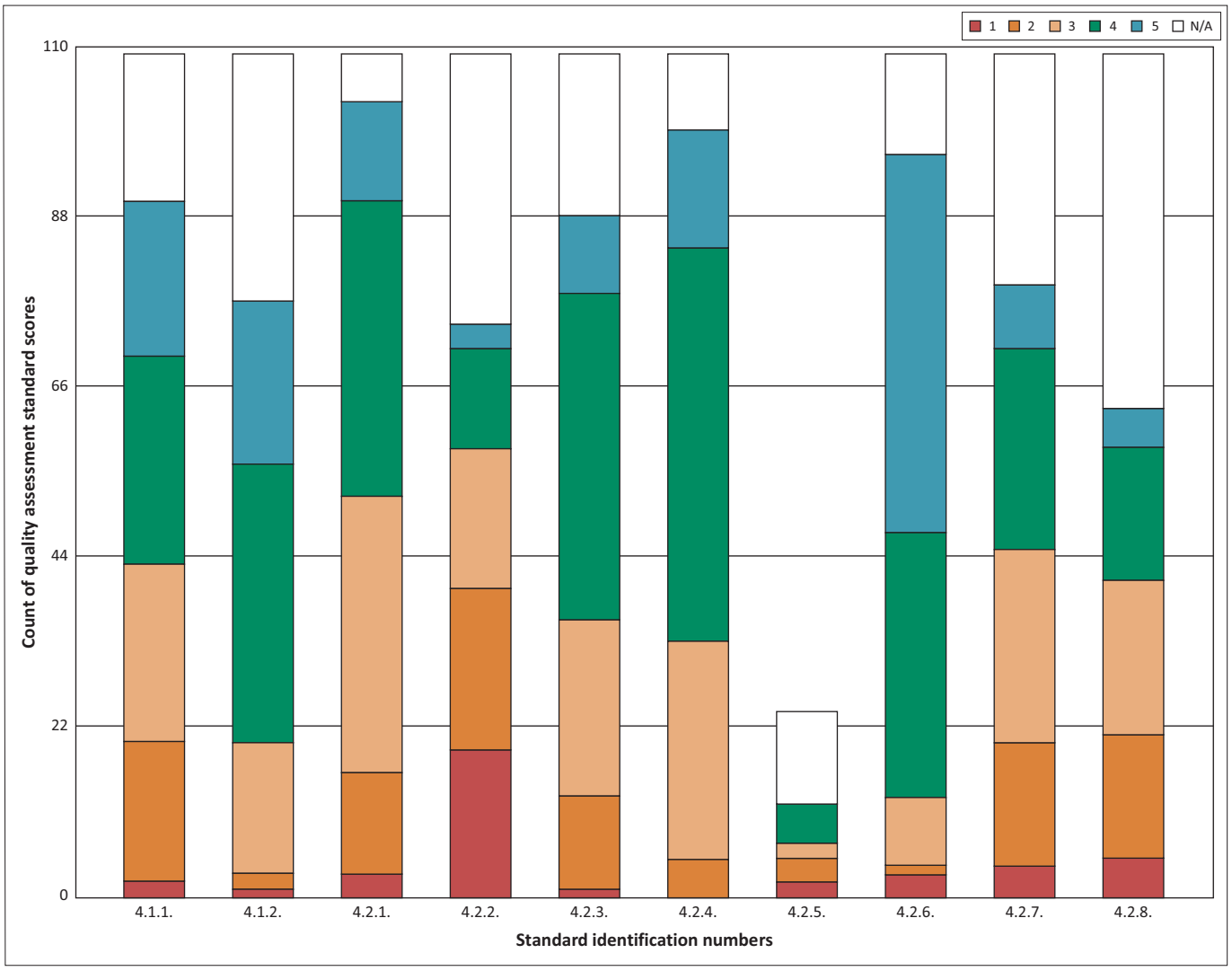

FIGURE 6: Distribution of ratings per standard: Follow-up, Use and Learning Phase.

such a short time of completing the evaluation, particularly for NEP evaluations that must go for quality assessment before being considered for approval by Cabinet. This has prompted further investigation of a more appropriate tool and approach for assessing utilisation at a later stage.

Scoring of evaluation quality by overarching consideration

The graph below presents average scores for six of the seven overarching considerations based on an aggregate measure of the standards aligned to that consideration. An average score for project management could not be produced at this time on account of the timing of its introduction after the start of the round 2 quality assessments.

Figure 7 highlights the areas where the sampled evaluations fared particularly poorly or were of a good standard. The most striking of these is the average score of 2.47 received for capacity development. Going back to the planning and design phase, and considering how the opportunity for learning was missed amongst standards there, the findings from this analysis are indicative of a broader trend within

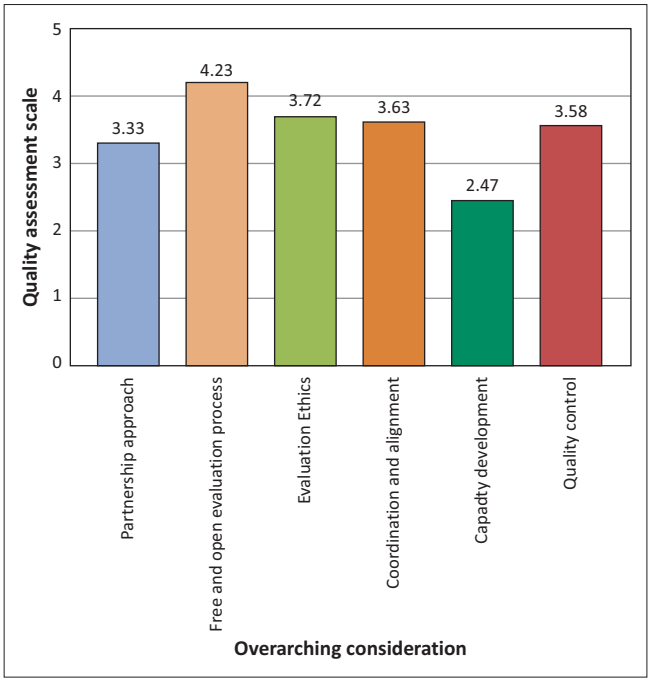

FIGURE 7: Average score of overarching considerations.

the sampled set of evaluations to neglect the capacity development opportunity presented by evaluation work. This is particularly concerning given the need to develop the skills and capacity of the public service with regard to evaluation practice.

The overarching consideration of a free and open evaluation process is rated highly at 4.23. However, closer scrutiny of the ratings and weighting system reveals that this consideration is, in particular, distorted as a result of the historical application of fewer standard items, including public access to an evaluation report, which biases it towards a higher rating. Again there is a self selection bias in that the evaluations in the sample were those which departments were willing to make available, and therefore more likely to be open about the evaluation process. Since adjusting the alignment standard items between rounds 2, there has been a decline of 0.21 points, in part because it is now more representative of a broader set of related standards.

National evaluation repository

Evaluation approval and publishing

Core to DPME's approach to evaluation is ensuring quality and an evaluation repository was developed to make available the findings of all evaluations that have been undertaken within national and provincial departments. This is aimed at facilitating informed decisions about programmes being implemented across all spheres of government. The repository can be accessed at: http://evaluations.dpme.gov.za/sites/EvaluationsHome/SitePages/Home.aspx

Conclusion

Evaluation practice

For the meta-evaluations conducted of the respective samples of 83 and 25 evaluations, the quality assessment found that evaluation practice is generally rated well although the second round saw a slight decline in quality, based on overall scores. However, this decline should be seen within the context of an expanded set of standards, as well as the fact that in round 2 much more information was available about the process, and not just the product, providing more evidence.

Implementation and impact evaluations still constitute the majority of evaluations undertaken to date, but there has been a drop in impact evaluations as a proportion of the overall total. Data from the second round suggest a shift towards more formative assessments, a large contributor being that data often do not exist for impact evaluations, as this was not built into programme design. This is a significant issue moving forward.

The body of all evaluations assessed has generally shown above satisfactory levels of methodological appropriateness for relevant standards in the implementation and reporting phases and, in particular, has employed appropriate data gathering techniques consistent with the type of evaluation and its objectives. Background reviews of legislative, policy or programme contexts together with literature reviews have generally been executed to an above adequate standard.

However, there are a number of specific shortcomings in evaluation practice that continue. Programme intervention logic or a theory of change is not explicitly referred to in either the evaluation design or in the drawing of conclusions in many cases. Clarity of the intervention logic or theory of change, and regular reference to this, is critical to good evaluation practice, hence DPME's recent requirement that this occurs as part of all NEP evaluations.

New shortcomings identified in round 2 include the lack of preparatory review, either by peers or in testing data collection instruments according to specific standards in the implementation phase. Further, analytical rigour and the lack of exploration of alternative interpretations of evaluation findings in deliberating on the conclusions are two shortcomings persisting across both rounds in the reporting phase, which require urgent attention through support and guidelines.

Round 2 quality assessments also posed problems for assessing *Phase 4: Follow-up, use and learning*, because of the short timeframe between completion and assessment of government evaluations.

When considering the overarching considerations, capacity development stands out as an area in need of concerted improvement in evaluation practice. The lack of planned and well executed capacity development is a particular concern which DPME is using as one of the criteria in assessing evaluation proposals. Unless addressed, this lack of capacity poses a risk to the state's capability to effectively oversee, manage and utilise evaluations in the future.

Quality assessment practice

The quality assessment practice, including the methodology and tools applied, has produced a useful analysis of the quality of evaluations to date. A quality assessment tool has now been developed with an electronic platform that can be applied to future evaluations. The system is currently being expanded to offer evaluation planning, management and a document repository function, in addition to quality assessment. The online platform would thereby facilitate an evaluation lifespan tracking and monitoring system, complete with quality assessment. A subsequent rapid utilisation assessment is also being investigated two years after the conclusion of the evaluation to revisit use, learning and improvement.

The benefit in undertaking quality assessments of evaluations is that it opens up the potential to learn, and improve evaluations, through improved standards and support so that they increasingly be on par with best practice. Between

the first two rounds these standards have expanded to cater for all elements of evaluation practice. Going forward, it is necessary to consolidate those standards most critical to determining evaluation quality and refine the tool to produce the most credible quality assessment possible.

Acknowledgements

In addition to the authors, we wish to acknowledge the team of quality assessors: Fatima Rawat, Kevin Foster, Meagan Jooste, Katie Gull, Tim Mosdell, Nana Davies, Cathy Chames, Wilma Wessels, Robin Richards, Raymond Basson, Stephen Rule, Thandeka Mhlantla, Chiweni Chimbwete, Lewis Ndlovu and Kevin Kelly. Sean Walsh assisted in the development of software systems. Rosina Maphalla and Nazreen Kola also contributed to an earlier article on the quality assessment system.

Competing interests

The authors declare that they have no financial or personal relationship(s) that may have inappropriately influenced them in writing this article.

Authors' contributions

This article is the scholarly culmination of various pieces of evaluation related work commissioned by the DPME. I.G., C.J. and M.E. contributed to the background, South African standards, and national evaluation repository sections. M.L. and N.M. wrote the sections on the quality assessment system and results. D.P. contributed to the background section and T.B. to the audit of government evaluations.

References

African Evaluation Association (AfrEA), 2006, *African evaluation guidelines: Standards and norms,* 2006/7 version, African Evaluation Society.

American Evaluation Association (AEA), n.d., *The programme evaluation standards: Summary form,* American Evaluation Association.

Canadian Evaluation Society (CES), 2012, *Program evaluation standards,* Canadian Evaluation Society.

DeGEval Evaluation Society, 2002, *Evaluation standards,* DeGEval Evaluation Society.

Development Assistance Committee (DAC), 2010, *Quality standards for development evaluation,* Organisation for Economic Co-operation and Development, Paris.

Department of Performance Monitoring and Evaluation (DPME), 2011, *National evaluation policy framework,* Department of Performance Monitoring and Evaluation, Pretoria.

Department of Performance Monitoring and Evaluation (DPME), 2012, *Evaluation competency framework for government,* Department of Performance Monitoring and Evaluation, Pretoria.

Department of Performance Monitoring and Evaluation (DPME), 2013a, 'Assessment of government evaluations: Overview report' (unpublished), Department of Performance Monitoring and Evaluation, Pretoria.

Department of Performance Monitoring and Evaluation (DPME), 2013b, DPME *evaluation guideline 2.2.1- How to develop a terms of reference for an evaluation project,* viewed 11 November 2014, from http://www.thepresidency-dpme.gov.za

Department of Performance Monitoring and Evaluation (DPME), 2014, *Standards for evaluations in government,* Department of Performance Monitoring and Evaluation, Pretoria.

Joint Committee on Standards for Educational Evaluations (JCSEE), 1994, *The program evaluation standards: How to assess evaluations of educational programs,* Sage, Newbury Park, CA.

King, J. & Podems, D., 2014, 'Introduction to professionalizing evaluation: A global perspective on evaluator competencies', *Special Edition of the Canadian Journal of Evaluation* 28(3).

Podems, D., 2012, 'Evaluation standards and competencies for the South African government', (unpublished), Department of Performance Monitoring and Evaluation, Pretoria.

Podems, D. & Podems, D.R., 2014, 'Evaluator competencies and professionalizing the field: Where are we now?', *Special Edition of the Canadian Journal of Evaluation* 28(3).

Tarsilla, M., 2014, 'Evaluation capacity development in Africa: Current landscape of international partners, initiatives, lessons learned and the way forward', *African Evaluation Journal* 2(1).

United Nations Evaluation Group (UNEF), 2005, *Standards for evaluation in the UN system,* United Nations Evaluation Group, New York.

Windmer, T., Landert, C. & Bachmann, N., 2000, *Evaluation standards of SEVAL,* Swiss Evaluation Society, Geneva.