unicef
Office of Research-Innocenti

Methodological Briefs
Impact Evaluation No. 4

Evaluative Reasoning

E. Jane Davidson

RMIT UNIVERSITY
3ie International Initiative for Impact Evaluation
BetterEvaluation
unicef Office of Research-Innocenti

UNICEF OFFICE OF RESEARCH

The Office of Research is UNICEF's dedicated research arm. Its prime objectives are to improve international understanding of issues relating to children's rights and to help facilitate full implementation of the Convention on the Rights of the Child across the world. The Office of Research aims to set out a comprehensive framework for research and knowledge within the organization, in support of UNICEF's global programmes and policies, and works with partners to make policies for children evidence-based. Publications produced by the Office are contributions to a global debate on children and child rights issues and include a wide range of opinions.

The views expressed are those of the authors and/or editors and are published in order to stimulate further dialogue on impact evaluation methods. They do not necessarily reflect the policies or views of UNICEF.

OFFICE OF RESEARCH METHODOLOGICAL BRIEFS

UNICEF Office of Research Methodological Briefs are intended to share contemporary research practice, methods, designs, and recommendations from renowned researchers and evaluators. The primary audience is UNICEF staff who conduct, commission or interpret research and evaluation findings to make decisions about programming, policy and advocacy.

This brief has undergone an internal peer review.

The text has not been edited to official publication standards and UNICEF accepts no responsibility for errors.

To consult and download the Methodological Briefs, please visit http://www.unicef-irc.org/KM/IE/

For readers wishing to cite this document we suggest the following form:
Davidson, E. J. (2014). Evaluative Reasoning, *Methodological Briefs: Impact Evaluation 4*, UNICEF Office of Research, Florence.

Acknowledgements: This brief benefited from the guidance of many individuals. The author and the Office of Research wish to thank everyone who contributed and in particular the following:

Contributors: Simon Hearn, Patricia Rogers, Jessica Sinclair Taylor
Reviewers: Nikola Balvin, Karin Heissler, Debra Jackson

September 2014

UNICEF Office of Research - Innocenti
Piazza SS. Annunziata, 12
50122 Florence, Italy
Tel: (+39) 055 20 330
Fax: (+39) 055 2033 220
florence@unicef.org
www.unicef-irc.org

1. EVALUATIVE REASONING: A BRIEF DESCRIPTION

Evaluation, by definition, answers *evaluative* questions, that is, questions about quality and value. This is what makes evaluation so much more useful and relevant than the mere measurement of indicators or summaries of observations and stories.

Decision makers frequently need evaluation to help them work out what to do to build on strengths and address weaknesses. To do so, they must know not only what the strengths and weaknesses are, but also which are the most important or serious, and how well or poorly the programme or policy is performing on them. For example, they need to know not only by how much a certain outcome[1] has shifted, but also the *quality and value* of this outcome.

To answer evaluative questions,[2] what is meant by 'quality' and 'value' must first be defined and then relevant evidence gathered. Quality refers to how good something is; value refers to how good it is in terms of the specific situation, in particular taking into account the resources used to produce it and the needs it was supposed to address. Evaluative reasoning is required to synthesize these elements to formulate defensible (i.e., well reasoned and well evidenced) answers to the evaluative questions.

Evaluative reasoning is a building block in evaluation: it is used throughout the evaluation to synthesize information necessary to draw evaluative conclusions. This is done in two ways, by combining:

- evidence about performance on a particular dimension and interpreting it relative to definitions of 'how good is good' to generate a rating of performance on that dimension
- ratings of performance on several dimensions to come to an overall conclusion about how good performance is for a particular site, project, programme, policy or other 'evaluand' (a generic term for that which is being evaluated).

Evaluative reasoning is a requirement of *all* evaluations – formative, summative or developmental; process, outcome or impact; qualitative, quantitative or mixed methods; and primary or secondary. It is also a requirement for *all* design options.

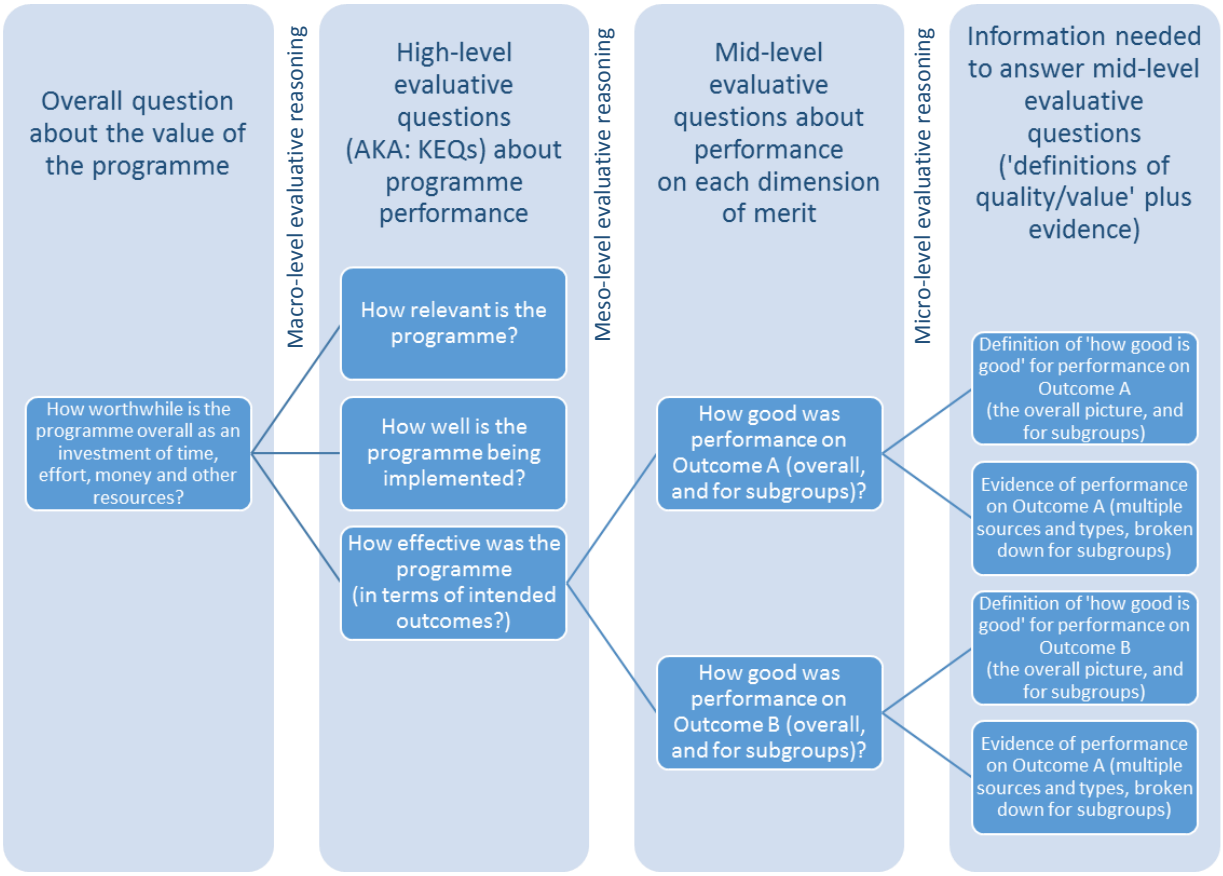
Figure 1 shows how the overall programme or policy value/worth is determined by answering a set of high-level or key evaluation questions (KEQs; see also Brief No. 3, Evaluative Criteria). An evaluation should have a limited set of high-level questions which are about performance overall. Each of these KEQs should be further unpacked by asking more detailed questions about performance on specific dimensions of merit (i.e., the KEQ is operationalized by several mid-level (meso) questions; depicted in Figure 1, reading from left to right) and sometimes even lower-level (micro) questions (not shown in Figure 1).

Evaluative reasoning is the process of synthesizing the answers to lower- and mid-level questions into defensible judgements that directly answer the high-level questions. All evaluations require micro- and meso-level evaluative reasoning (i.e., synthesizing answers from right to left in Figure 1); not all require it at the macro level.

[1] An 'outcome' is *something that happens* to people, communities, the environment or the economy, which is at least partially caused by the programme or policy. Outcomes may be positive or negative, intended or unintended, change that happen or changes that are prevented from happening. In the context of evaluation, use of the term outcome implies that the programme/policy contributed to, helped to cause and/or catalyzed (speeded up) the change.

[2] For an overview of the different types of questions (descriptive, causal, evaluative) that evaluations can address, see Brief No. 1, Overview of Impact Evaluation.

Figure 1. Levels of evaluative reasoning to determine the overall value/worth of an intervention



Evaluative reasoning must begin by identifying the evaluative criteria (see Brief No. 3, Evaluative Criteria) and the evidence that will be used to assess performance, and how these will be synthesized and used to produce well reasoned and well evidenced answers to each evaluative question.

Box 1 contains an example taken from an evaluation of an educational leadership programme, which demonstrates an executive summary that explains the key findings in an explicitly evaluative way. It includes a clear statement about the quality of the programme and a sense of what the main weaknesses were, and – importantly – how serious these were. There is also a brief synopsis of the basis on which these conclusions were drawn, and the reader is referred to a clear breakdown of findings in the easy-to-read 'snapshot' bar graph in table 1 that uses *explicitly evaluative ratings* ('excellent', 'adequate', etc.) for each criterion to make it clear how good performance was for each aspect.

Box 1. Example of key findings explained in an explicitly evaluative way

Key findings from the study

Overall, the review found that the First-time Principals Induction Programme (FTPIP) offered a good quality induction programme for first-time principals (FTP). There were some aspects of the programme which required fine-tuning to maximize their effectiveness, but no major changes were recommended by the review. The review found the FTPIP provided a good platform for inducting first-time principals across primary, secondary and (to a lesser extent) kura kaupapa settings to be educational leaders in their schools.

The FTP programme was a fantastic initiative as far as I am concerned and money well invested. I felt supported and had connections to contact when I needed to do so. Our mentors and the PLG [Professional Learning Group] members could share experiences and give timely advice. Sessions at each of the residential courses were scarily apt and timely. Thank you for such a great programme. It was a combination of contextual, real life learning that could be shared with/by experienced people/principals. (FTP participant)

The following Table 1 provides an executive snapshot, identifying the evaluative criteria on which performance was strongest and weakest as judged by FTPs, stakeholders and sector leaders.

Overall FTPs believed they were mostly (with some exceptions) better prepared to be a leader of learning and able to apply this learning in their schools, compared to before they started the FTPIP.

Source: Reproduced under a Creative Commons licence from: Oakden, Judy, *Evaluation rubrics: how to ensure transparent and clear assessment that respects diverse lines of evidence,* BetterEvaluation, 2013, p. 12. See http://betterevaluation.org/resource/example/rubrics-oakden.

Table 1. Example of a clear breakdown of findings that uses explicitly evaluative ratings (accompanies box 1)

Table 1: Summary of evidence that FTPIP is of good quality–FTPs as 'leaders of learning'

Evaluative Criteria		Ratings				
		Poor	Adequate	Good	Very good	Excellent
Principals participate in professional learning and are recognised as 'leading learners' in their school	Overall rating					
	There are clear and appropriate professional development goals set for FTPs					
	Good working relationships which provide professional support and advice to FTPs are established between FTPs and mentors					
	Networks of peer support are established between FTPs and peers					
	FTPs engage in reflective learning about being leading learners in their schools					
	FTPs understand the importance of being leaders of learning and have clear strategies to effect this					
	FTPs report they know how to collect, analyse and act on data to support student learning					
	There is evidence of the FTPs focus on equity for Māori					
	Support for FTPs is well co-ordinated (especially where there are several support groups working with the FTP)					

Source: Ibid.

Main points

1. Evaluative reasoning is used throughout the evaluation to systematically combine evidence together with definitions of 'quality' and 'value' to come to defensible conclusions about performance.
2. Evaluative reasoning is a requirement of all evaluations, irrespective of purpose, method or design.
3. Evaluative reasoning is the opposite of 'operationalization'. When an evaluative question is operationalized, it is broken down to reveal what evidence to gather and what definitions of quality and value to apply. Evaluative reasoning is the process of packing those elements back together to answer the evaluative question.

2. WHEN IS IT APPROPRIATE TO USE THIS METHOD?

All evaluations require evaluative reasoning. The Terms of Reference (ToR) for an evaluation should specifically ask how the evaluative reasoning will be done. By setting clear expectations for this, managers can help to ensure that the evaluation delivers direct and explicitly evaluative answers to the KEQs – and a conclusion about the overall quality and value of the programme or policy, if required.

To judge the quality and value of its programmes and policies, UNICEF adopts the standard OECD-DAC evaluation criteria:[3] relevance, effectiveness, efficiency, impact and sustainability (see Brief No. 3, Evaluative Criteria). UNICEF evaluations must also determine the extent to which equity, gender equality and human rights[4] are reflected in the intervention design, its implementation and the results (for humanitarian assistance interventions, additional evaluative criteria may be appropriate).[5] This means that evaluations should include a range of lower-level evaluation questions to address these specific dimensions. The evaluative criteria should also be considered in how 'success' is defined and used in the evaluative reasoning process.

Evaluation, by definition, must answer truly evaluative questions: it must ask not only 'What were the results?' (a descriptive question) but also '*How good* were the results?'. This cannot be done without using evaluative reasoning to evaluate the evidence relative to the definitions of quality and value.

Micro-level evaluative reasoning is required in *all* evaluations. Micro-level evaluative reasoning consists of combining definitions of quality and value with evidence in order to draw conclusions about how good was the performance on key dimensions (such as outcomes).

Without micro-level evaluative reasoning, evaluation reports merely summarize evidence without drawing evaluative conclusions or directly answering evaluative questions. Reporting results against established targets or metrics still falls short of drawing an explicitly evaluative conclusion. Even when using goal-based evaluation part of the core task is to determine whether 'met target' is in fact the appropriate criterion for concluding 'good result'. Furthermore, if the result either falls short of the target or clears it substantially, it is necessary to be able to say whether this is 'completely unacceptable', 'barely acceptable', 'very good'

[3] Organisation for Economic Co-operation and Development – Development Assistance Committee, 'DAC Criteria for Evaluating Development Assistance', web page, OECD, http://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm.

[4] Universalia Management Group, *Global Evaluation of the Application of a Human Rights Based Approach in UNICEF programming*, *Final Report – Volume I*, UNICEF, 2012.

[5] United Nations Children's Fund, 'Linking evaluation criteria to evaluation questions: Additional criteria for humanitarian assistance programmes', in the UNICEF Monitoring and Evaluation Training Resource, UNICEF. (At the time of writing, this resource is currently under revision.)

or 'excellent', and how these categories are defined. Simply pointing out where performance falls relative to targets is insufficient.

In some cases, evaluative conclusions are drawn (e.g., a claim is made that something is 'a good outcome') but the evaluative reasoning is not explicit. When this happens, evaluation appears as if based on personal opinion rather than sound evaluative reasoning; the reasoning must be made explicit to demonstrate that evaluative conclusions are justified (examples of how to do this are provided later in this brief). In addition to simply being good professional practice, clear and transparent evaluative reasoning allows others to follow the logic and point out any flaws, which in turn allows for the quality of the work to be strengthened.

Meso-level evaluative reasoning is required to pack together evidence on multiple dimensions (e.g., all intervention outcomes or all aspects of design) to answer KEQs (such as 'How good are the outcomes?' or 'How well designed is the intervention?').

Many, if not most, evaluations tend to leave the evidence disaggregated. This is problematic for several reasons.

Firstly, *all* interventions will have a mix of strengths and weaknesses, success stories and disappointments. For the manager trying to make decisions about what to do next, it is vital to know which of the weaknesses are the most serious and should therefore be addressed first and fast. In other words, the 'importance' component of evaluative reasoning is critical for making evaluation actionable.

Secondly, even when an evaluation is done for formative purposes (i.e., to improve or reorient a programme or policy), it is almost invariably important to know, for example, whether the outcomes (as a set, not just individually) are strong enough given where the programme/policy is in its development. There is usually at least one stakeholder (e.g., a funder) who needs to know sooner rather than later that the investment seems to be on track to deliver enough in the way of valuable outcomes.

Thirdly, it is common to see 'answers' to KEQs generated not by synthesizing the intervention's performance across all of the criteria but instead by choosing the one or two metrics that are deemed most important and presenting these as if they are 'the answer'. While it is valid to *weight* (qualitatively or quantitatively) the more important considerations more heavily, it is not valid to privilege a small number of metrics to the *exclusion* of other dimensions of merit.

Macro-level evaluative reasoning is required whenever there is an overarching question about the value or worth of the entire programme or policy. This overarching question is answered by synthesizing answers to all of the KEQs, including those around intervention design, implementation, intended and unintended outcomes, and comparative cost-effectiveness.

To summarize, for *impact* evaluations:

- evaluative reasoning is *always* needed at the micro level – to determine how good the results are on particular dimensions
- evaluative reasoning is always needed at the meso level – to adequately answer the KEQs the evaluation has set out to answer in the first place
- evaluative reasoning is often (but not always) needed at the macro level – to come to an overall conclusion about whether the intervention is a worthwhile investment of time, effort, money and other resources.

3. HOW TO DO EVALUATIVE REASONING

Evaluative reasoning is most easily done in evaluations where it is built into the design from the beginning. It is possible to do it retrospectively, but this is more difficult. If needs or priorities change, or if outcomes are emergent or unanticipated, evaluative reasoning must be done in 'real time'.

The following steps trace evaluative reasoning from the evaluation design phase, but with options outlined for those times when evaluative reasoning is a late addition to the evaluation and has to be retrofitted to data that have already been gathered. All of the following may be done either collaboratively (with stakeholder involvement) or independently (by the evaluation team).

1. **Develop a set of KEQs to guide the whole evaluation.** This should be undertaken during evaluation scoping and planning (ideally based on a draft in the ToR). Based on the author's experience, 7 ± 2 questions is a good number in general. This allows for coverage of different aspects of the programme/policy, but is a small enough number of questions that the reader will not be overwhelmed. Each of these high-level questions will be broken down into a set of sub-questions.

2. **Identify the appropriate basis for defining what a 'valuable outcome' and a 'good quality programme/policy' should look like in this context**. Specifically, consider dimensions related to the evaluative criteria mentioned above (e.g., OECD-DAC; human rights based approach to programming or HRBAP). Key sources here will be strengths and needs assessments, community values and aspirations, relevant standards for the particular type of initiative and so on.

   Programme/policy planning documents with objectives and key strategies will also be useful here. To be able to say whether an outcome is 'good', however, it is necessary to find a stronger basis than simply defining 'good' as 'met target' – and it is also necessary to be able to say, for example, how good a near miss on the target would be.

3. **Devise one or more evaluative rubrics to define what the *mix* of evidence will look like at different levels of performance**. An evaluative rubric is a table that describes what the evidence should look like at different levels of performance, on some criterion of interest or for the programme/policy overall.

   The key here is to think not in terms of simply creating groupings on a quantitative indicator or a tallied score (e.g., 90 to 100 per cent = excellent) but rather to discuss what the *mix of evidence* (qualitative, quantitative or both) should look like at each level of performance.

   Occasionally it may be necessary to create a rubric based on a single and very important measure. Even in such cases, the key is to consider what the evidence will look like not only overall but also for particular subgroups (e.g., it is often important to give special consideration to outcomes for vulnerable populations) for an outcome to be judged 'good' (i.e., worthwhile, equitable, most needed, etc.). In most cases, the infusion of qualitative evidence will also help to put 'flesh' on those quantitative 'bones', giving a clearer sense of what the outcome looks like in reality.

4. **Design data collection instruments and protocols to gather evidence that is most central to the definition of quality, as defined in the evaluative rubric**. The rubric paints the picture of what the mix of qualitative and quantitative evidence would look like, and this also gives a clear sense of what will be needed to determine how performance should be rated.

   In evaluations where existing data need to be used or the evidence has already been gathered, the key is not to write the rubric solely around what is available, but rather to paint the broad picture of what performance looks like regardless of what evidence is available.

5. **Conduct micro-level evaluative reasoning**. Interpret evidence relative to definitions of quality and value, using the evaluative rubric or another appropriate evaluative interpretation guide.

6. **Use needs assessment and other relevant evidence to determine which considerations are critically important, which are important but not key, and which are beneficial to have but not that important**. Like all steps in this process, this may be done collaboratively with stakeholder involvement or independently by the evaluation team.

7. **Based on the above assessment of importance, determine how the evidence on multiple dimensions should be synthesized to generate answers to the KEQs**. This is not simply a case of assigning scores and weights (as in a 'numerical weight and sum' approach). Meso-level

evaluative reasoning must be judgement-based so that it can incorporate 'hurdles'. Hurdles are minimum levels of performance on certain key dimensions that are required for the programme/policy to be considered even minimally acceptable on that evaluation question (a 'hard hurdle' or 'bar') or for the programme/policy to achieve a rating at a certain level or above (a 'soft hurdle')[6].

8. **Have the evaluation team, drawing on relevant expertise as appropriate, consider the evidence and perform the meso-level evaluative reasoning**. The end product of this should be direct and explicitly evaluative answers to each of the KEQs used to frame the entire evaluation. These direct answers are what should be summarized in the executive summary of the report.
9. **If macro-level evaluative reasoning is also required, generate and apply a whole programme/policy rubric that describes what programme/policy performance will look like overall, taking into account the answers to all of the KEQs**. Again, this should not involve a numerical weight and sum exercise, but rather the painting of a picture of what the array of performances will look like if the programme/policy is to be judged hugely successful and valuable (as opposed to barely adequate, or a 'failure').

Where possible, it is highly advisable that rubrics for all three levels of evaluative reasoning are drafted *before* the evidence is gathered. This enables clearer thinking and discussion about how success should be defined without the temptation to cast available evidence in a positive light.

Involving stakeholders in the development of evaluative rubrics not only improves validity and credibility, but also allows decision makers to mentally 'rehearse' what the evidence might look like and understand ahead of time how it will be interpreted. This reduces misunderstandings and disagreements about the interpretation of evidence later on, thereby improving the acceptance of the findings as valid and credible as well as the commitment to taking any necessary action.

Evaluative rubrics should always be treated as living tools, and should be used for **emergent outcomes and shifting priorities** as well as those that are anticipated from the outset.

In cases where needs or priorities have changed, rubrics should be revised to reflect this. In some cases, this will require some capturing of different evidence to better understand how well the programme/policy has performed.

For emergent and unanticipated outcomes – just as with intended and anticipated outcomes – it is still necessary to be able to say how beneficial or detrimental these outcomes were relative to the needs and aspirations of the community. In such cases, rubrics must be developed as soon as emergent outcomes are identified, but ideally before all of the evidence is gathered. This allows for the real-time streamlining of evidence capture and provides a workable frame for interpreting that evidence when it comes in.

4. PRACTICAL CONSIDERATIONS

The primary considerations when planning for evaluative reasoning are:

- What are appropriate ways to define what quality and value mean in this context, for this programme or policy, and why? Whose voices should be at the table for this?
- Have the data collection instruments already been designed, so that the evaluative reasoning will be done with what is already available? Or is it early enough to allow the evaluative reasoning to inform the design of the data collection instruments?

[6] Davidson, E. Jane, Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation, Sage, Thousand Oaks, 2005.

- Who should be involved in the evaluative reasoning process itself? Key considerations are:
    - **Validity** – Whose expertise is needed to get the evaluative reasoning right? As well as a lead evaluator with expertise in evaluation-specific methodology, it will be necessary to include people with expertise in the subject matter; in the local context and culture; and in interpreting the particular kinds of evidence to be used.
    - **Credibility** – Who must be involved in the evaluative reasoning to ensure that the findings are believable in the eyes of others?
    - **Utility** – Who is most likely to use the evaluation findings (i.e., the products of the evaluative synthesis)? It may be helpful to have primary intended users involved so that they can be part of the analysis and see the results with their own eyes. This increases both understanding of the findings and commitment to using them.
    - **Voice** – Who has the right to be at the table when the evaluative reasoning is done? This is particularly important when working with groups of people who have historically been excluded from the evaluation table such as indigenous peoples. Should programme/policy recipients or community representatives be included? What about funders? And programme managers?
    - **Cost** – Taking into consideration the opportunity costs of taking staff away from their work to be involved in the evaluation, at which stages of the evaluative reasoning process is it best to involve different stakeholders for the greatest mutual benefit?

A major challenge is that evaluative reasoning, although essential, is not yet a common skill set in the repertoire of evaluation professionals, as it is absent from the curricula of virtually all advanced evaluation training and graduate programmes. Unsurprisingly, it is also not yet widely understood among those who commission and oversee evaluation work.

A second major challenge is that part of the task of defining what constitutes a 'good' intervention involves considering whether or not the intervention itself is ethical. This is another area in which evaluation training is sorely lacking, and most evaluation teams will be ill-equipped to conduct ethical analyses of programmes/policies themselves.

Managers seeking to contract evaluation teams with these skill sets may require support to clearly set out evaluative reasoning expectations in the ToR – including any need for an ethical analysis of the intervention itself. Managers may also require access to expert advice and support throughout the evaluation commissioning and management processes, so that they can negotiate and manage effectively what is needed, even with contractors who may themselves still be mastering this methodology.

5. WHICH OTHER METHODS WORK WELL WITH THIS ONE?

Evaluative reasoning is a requirement of all evaluations, irrespective of the methods or evaluation approach used. The bottom line is that *all* evaluations are tasked with asking and answering evaluative questions, not simply with measuring and describing results. All findings must therefore be interpreted within an evaluative frame. This means not only saying what the results are, but also how good they are. Evaluative reasoning is necessary to be able to do this.

6. PRESENTATION OF RESULTS AND ANALYSIS

The structure of an evaluation report can do a great deal to encourage the succinct reporting of direct answers to evaluative questions, backed up by enough detail about the evaluative reasoning and methodology to allow the reader to follow the logic and clearly see the evidence base.

The following recommendations will help to set clear expectations for evaluation reports that are strong on evaluative reasoning:

1. The executive summary must contain direct and explicitly evaluative answers to the KEQs used to guide the whole evaluation.
2. Explicitly evaluative language must be used when presenting findings (rather than value-neutral language that merely describes findings). Examples should be provided (e.g., of the type seen in the good practices example, below).
3. Use of clear and simple data visualization to present easy-to-understand 'snapshots' of how the intervention has performed on the various dimensions of merit.
4. Structuring of the findings section using KEQs as subheadings (rather than types and sources of evidence, as is frequently done).
5. There must be clarity and transparency about the evaluative reasoning used, with the explanations clearly understandable to both non-evaluators and readers without deep content expertise in the subject matter. These explanations should be broad and brief in the main body of the report, with more detail available in annexes.
6. If evaluative rubrics are relatively small in size, these should be included in the main body of the report. If they are large, a brief summary of at least one or two should be included in the main body of the report, with all rubrics included in full in an annex.

A hallmark of great evaluative reasoning is how succinctly and clearly key points can be conveyed without glossing over important details.

7. EXAMPLES OF GOOD PRACTICES

Because few development impact evaluations systematically and explicitly use evaluative reasoning to answer evaluative questions, the example in box 2 refers instead to an educational evaluation of a school leadership programme.

This case study illustrates the use of KEQs and evaluative rubrics for evidence interpretation. Each high-level question is explicitly evaluative, asking not only 'what' or 'how' but also 'how good'.

Box 2. Good practice case study: Evaluation of the First-Time Principals Induction Programme

This evaluation was framed around four KEQs, the main one being:

> *To what extent is the First-time Principals Induction Programme a high-quality programme that inducts first-time principals across primary, secondary and kura kaupapa (indigenous Māori language immersion) settings to be educational leaders in their schools?*

Rubrics were constructed collaboratively, with particular care paid to ensuring that there was a Māori (indigenous) perspective present throughout the evaluative reasoning process. This was important for validity, credibility, justice and utilization reasons. A sample of the top two levels of one of the rubrics is shown below. Of particular note is the incorporation of different aspects of performance within the rubric as well as a clear global statement at the top of each rating category to explain the general intent.

Rubric for: Participate in professional learning and are recognized as 'leading learners' in their school

Rating	Evaluative criteria
Excellent	Clear example of exemplary performance or best practice in this domain: no weaknesses • There are always clear and appropriate professional development goals set for first-time principals • Good working relationships which provide professional support and advice to first-time principals are always established between first-time principals and mentors • Networks of peer support are always established between first-time principals and peers • First-time principals always engage in reflective learning about being leading learners in their schools • First-time principals always understand the importance of being leaders of learning and have clear strategies to effect this • First-time principals always report they know how to collect, analyse and act on data to support student learning • There is always evidence of the first-time principal's focus on equity for Māori • Support for first-time principals is well coordinated (especially where there are several support groups working with the first-time principal).

Very good	Very good or excellent performance on virtually all aspects; strong overall but not exemplary; no weaknesses of any real consequence • There are almost always clear and appropriate professional development goals set for first-time principals • Good working relationships which provide professional support and advice to first-time principals are almost always established between first-time principals and mentors • Networks of peer support are almost always established between first-time principals and peers • First-time principals almost always engage in reflective learning about being leading learners in their schools • First-time principals almost always understand the importance of being leaders of learning and almost always have strategies to effect this • First-time principals almost always report they know how to collect, analyse and act on data to support student learning • There is almost always evidence of the first-time principal's focus on equity for Māori • Support for first-time principals is almost always well coordinated (especially where there are several support groups working with the first-time principal).

Source: Reproduced under a Creative Commons licence from: Oakden, Judy, *Evaluation rubrics: how to ensure transparent and clear assessment that respects diverse lines of evidence,* BetterEvaluation, 2013, p. 12. See http://betterevaluation.org/resource/example/rubrics-oakden.

Good rubrics like these encourage the use of sound evaluative judgement, providing a shared language for interpreting evidence that increases both inter-rater reliability and validity. This approach contrasts with many approaches to 'quantification' that try to eliminate judgement as if it is inherently unreliable and based solely on opinion.

The evaluation also engaged in other good practices, including a multi-stage process for assessing data against the rubrics. All of the qualitative data (i.e., from interviews with key stakeholders) and quantitative data (i.e., from a survey with principals) were converted to ratings – on a scale from excellent to poor – against each rubric, as depicted in table 2. This synthesis process allowed for overall evaluative judgements to be made about the results using all sources of data.[7]

[7] For details about this process, see Oakden, Judy, *Evaluation rubrics: How to ensure transparent and clear assessment that respects diverse lines of evidence,* Better Evaluation, 2013, p. 10. Available online: http://betterevaluation.org/sites/default/files/Evaluation%20rubrics.pdf.

Table 2. Ratings used to assess quantitative and qualitative data against each rubric

Rating	Quantitative and qualitative data
Excellent: Always	Clear example of exemplary performance or best practice in this domain; no weaknesses. Likely that 90 per cent or more agree with statement to a considerable or high degree.
Very good: Almost always	Very good to excellent performance on virtually all aspects; scoring overall but not exemplary; no weaknesses of any real consequence. Possibly 80 to 90 per cent agree with statement to a considerable or high degree.
Good: Mostly, with some exceptions	Reasonably good performance overall; might have a few slight weaknesses but nothing serious. In the range of 60 to 80 per cent agree with statement to a considerable or high degree, and no more than 15 per cent agree to a limited or very limited degree.
Adequate: Sometimes, with quite a few exceptions	Fair performance, some serious but non-fatal weaknesses on a few aspects. Around 40 to 60 per cent agree with statement to a considerable or high degree, and no more than 15 per cent agree to a limited or very limited degree.
Poor: Never (or occasionally, with clear weaknesses evident)	Clear evidence of unsatisfactory functioning; serious weaknesses across the board on crucial aspects. Probably less than 40 per cent agree with statement to a considerable or high degree.

Source: Reproduced under a Creative Commons licence from: Oakden, J. (2013). *Evaluation rubrics: How to ensure transparent and clear assessment that respects diverse lines of evidence.* Better Evaluation, p. 10. Available online: http://betterevaluation.org/sites/default/files/Evaluation%20rubrics.pdf; Oakden adapted the table content from Davidson, E. Jane, *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*, Sage, Thousand Oaks, 2005.

8. EXAMPLES OF CHALLENGES

Some evaluations fail to adequately use evaluative reasoning and instead present non-evaluative findings in a descriptive manner, with no clear conclusions about how good performance has been overall.

To illustrate some of the problems of failing to adequately use evaluative reasoning, an evaluation of a UNICEF-funded youth development programme is used as an example. The (slightly altered) excerpt from the executive summary shows non-evaluative facts presented with no comment regarding their merit or value:

> Under this scheme 14,000 youths received money as start-up capital, to start their own business. Of these 89 per cent of those who had started their own business were running it successfully. A higher proportion of girls were successful in running their business than boys. Among those whose businesses had been unsuccessful, the major reasons given were the high price of inputs and a lack of market for their products. Of those who took credit and started their own business 76 per cent said their lives had changed for the better as a result of their starting a small business. In terms of provision of start-up capital the cost per head for the financial support ranged from US$45 to US$250. In most regions for cost per head of the financial support showed no or a small increase over time. This can be related to the rising cost of inputs across the country in this time period.

> Both in terms of training and provision of capital the number of male beneficiaries was more than twice that of female beneficiaries.

The executive summary provides facts and figures about the number of participants, a comparison of participation rates and successful outcomes for males and females, the percentage of participants who were successful and the cost per head, without stating the quality or value of such results.

To a reader looking for evaluative conclusions from this report, the following questions remain unanswered:

- Was 14,000 youths a 'good' number to have supported with start-up capital, given the size of the investment for this component of the programme?
- What about the gender mix – with more than twice as many males as females benefiting – particularly given the fact that girls were more likely to be successful in business than boys?
- How good is 89 per cent of participants running their own business at the particular point in time assessed?
- Did this component of the programme deliver worthwhile value for money for the cost per head of financial support?

When results are presented in a non-evaluative manner, as in this example, the quality or value of the programme/policy is not at all evident in the executive summary. The reader instead has to apply personal judgements to the evidence to make sense of the findings.

The lack of evaluative reasoning also makes it difficult to see the logical link between these numbers and descriptions and the list of recommendations provided at the front of the report.

9. KEY READINGS AND LINKS

BetterEvaluation, 'Synthesise data from a single evaluation', web page, BetterEvaluation, http://betterevaluation.org/plan/synthesize_value/synthesize_data_single_evaluation. (Includes a downloadable, two-page PDF summarizing the main options for evaluative synthesis.)

The Center for Advanced Research on Language Acquisition, 'Process: Creating rubrics', web page, CARLA, 2012, http://www.icyte.com/saved/www.carla.umn.edu/560426.

Coyle, J., et al., 'Evaluation rubrics with E. Jane Davidson', *Adventures in Evaluation* podcast series, 2013, http://adventuresinevaluation.podbean.com/2013/12/08/evaluation-rubrics-with-e-jane-davidson/.

Dart, Jessica, et al., *Review of Evaluation in Agricultural Extension,* Rural Industries Research and Development Corporation, RIRDC Publication No. 98/136, 1998.

Davidson, E. Jane, Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation, Sage, Thousand Oaks, 2005.

Davidson, E. Jane, *Actionable evaluation basics: Getting succinct answers to the most important questions*, minibook, Real Evaluation, 2013. Available as a paperback, PDF or ebook, in English, Spanish and French. See http://realevaluation.com/read/minibooks/.

Evalue Research, *Final Evaluation Report of the Recognised Seasonal Employer Policy (2007–2009),* Department of Labour, New Zealand Government, Wellington, 2010. See http://dol.govt.nz/publications/research/rse-evaluation-final-report/rse-final-evaluation.pdf.

Oakden, Judy, 'Evaluation rubrics: how to ensure transparent and clear assessment that respects diverse lines of evidence', web page, BetterEvaluation, 2013, http://betterevaluation.org/resource/example/rubrics-oakden.

Parcell, Geoff, and Chris Collison, *No More Consultants: We know more than we think*, John Wiley & Sons, London, 2009. (Useful 'river diagram' is summarized at http://betterevaluation.org/resources/tools/rubrics/river_chart)

Rogers, Patricia J., and E. Jane Davidson, various blog articles on evaluative rubrics, Genuine Evaluation, http://genuineevaluation.com/?s=rubrics.

Scriven, Michael, *Evaluation Thesaurus*, fourth edition, Sage, Newbury Park, 1991. (Key source for understanding the underlying logic of evaluative synthesis.)

GLOSSARY

Developmental evaluation	*An evaluation approach which aims to understand the activities of a programme operating in a complex and dynamic environment, characterised by methodological flexibility and systems thinking. The focus is on innovation and strategic learning rather than standard outcomes. It facilitates real-time feedback to programme stakeholders and promotes continuous development, re-adjustment, and adaptation.*
Evaluative criteria	*The values (i.e. principles, attributes or qualities held to be intrinsically good, desirable, important and of general worth) which will be used in an evaluation to judge the merit of an intervention. Examples include OECD-DAC criteria, HRBAP and humanitarian assistance criteria.*
Evaluative questions	*Evaluative questions ask about the overall conclusion as to whether a programme or policy can be considered a success, an improvement or the best option.*
Evaluative rubric	*A table that describes what the evidence should look like at different levels of performance, on some criterion of interest or for the intervention (programme/policy) overall.*
Formative evaluation	*Evaluation intended to improve performance, most often conducted during the implementation phase of projects or programmes. Formative evaluations may also be conducted for other reasons such as compliance, legal requirements or as part of a larger evaluation initiative. See: Process evaluation.*
Impact	*Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended. (OECD-DAC definition, 2010)*
Impact evaluation	*An evaluation that provides information about the impacts produced by an intervention. It can be undertaken of a programme or a policy, or upstream work – such as capacity building, policy advocacy and support for an enabling environment. It goes beyond looking only at goals and objectives to also examine unintended impacts. See: impact.*
Key evaluation questions (KEQs)	*High-level (macro level) evaluation questions about overall performance, which the evaluation should aim to answer. KEQs are derived from the purpose of the evaluation.*
Mixed methods	*Research methods which collect and/or analyse both quantitative and qualitative data to study the same research question. Mixed methods allow the triangulation of data and more than one type of investigative perspective. See: qualitative research, quantitative research.*
Outcome	*The likely or achieved short-term and medium-term effects of a programme or policy's outputs, such as a change in vaccination levels or key behaviours.*
Primary research	*Research that collects and analyses new data collected for the purposes of the particular project. Examples include surveys, interviews and observations. See: secondary research.*
Process evaluation	*An evaluation of the internal dynamics of implementing organizations, their policy instruments, their service delivery mechanisms, their management practices, and the linkages among these. (OECD-DAC definition, 2010)* *See: formative evaluation.*

Reliability (inter-rater)	*Measure of reliability which assesses the degree to which different raters/judges/coders agree on their assessment decisions/ratings. It provides a score of how much consensus there is in the ratings of the raters. Related terms: inter-rater agreement.*
Qualitative data	*Descriptive data which can be observed, but not measured. It can include text, images, sound, etc. but not numeric/quantitative values. See: quantitative data.*
Qualitative research	*Research which collects and analyses qualitative data. Typical research methods include case study, observation, and ethnography. Results are not usually considered generalizable, but are often transferable. See: qualitative data, quantitative data, quantitative research.*
Quantitative data	*Measures of values or counts expressed as numbers. Quantitative data can be quantified and verified and used for statistical analysis. Results can often be generalized, though this is not always the case. See: qualitative data.*
Quantitative research	*Research which collects and/or analyses quantitative data. See: quantitative data, qualitative data, qualitative research.*
Secondary research	*Analysis of existing data, which were not collected primarily for the purposes of the given research project/evaluation. The data may have been collected as part of another research project/evaluation, a census, performance monitoring or something else.* *See: primary research.*
Summative evaluation	*Type of evaluation conducted at the end of an intervention (or a phase of that intervention) to determine the extent to which anticipated outcomes were produced. Summative evaluation is intended to provide information about the worth of a programme.*
Terms of Reference (ToR)	*A statement of the background, objectives, intended users, key evaluation questions, methodology, roles and responsibilities, timelines, deliverables, quality standards, evaluation team qualifications and other relevant issues which specify the basis of a UNICEF contract with the evaluators.*
Validity	*The degree to which a study accurately reflects or assesses the specific concept that the researcher is attempting to measure. A method can be reliable, consistently measuring the same thing, but not valid. See: reliability.*