

The overall evaluative conclusion as an element of evaluation practice

Jo MacDonald

The questions of who values, with whom, in what ways, and under what conditions concern all evaluators but are explicitly considered by some theorists more than others. Theorists placed on the valuing branch of Christie and Alkin's (2013) evaluation theory tree emphasise valuing in their conceptualisation of evaluation, but even among these theorists there is diversity in the ways in which valuing is considered and realised in evaluation practice. This article explores this diversity within one aspect of valuing—the valuing involved in reaching a warranted conclusion about the overall merit, worth, or significance of an evaluand. It considers the extent to which the literature discusses overall evaluative conclusions as an element of evaluation practice; the extent to which drawing such conclusions is seen as the responsibility of the evaluator or stakeholders; and the methods that may be used to reach a warranted evaluative conclusion. The author concludes that there has been little empirical research undertaken on the valuing involved in reaching a warranted conclusion about the overall merit, worth, or significance of an evaluand. Much of the literature is evaluators theorising from different epistemological positions. Thus, while the literature does

not definitively inform evaluators of whether they should always reach an overall evaluative conclusion, who they should involve, and what methods they should use, this review does support evaluators to reflect on these issues in their practice, and to make deliberate, informed decisions about the making—or not—of overall conclusions or judgements in future evaluations.

The questions of who values, how, and under what conditions are more explicitly considered by some theorists than others, and are contentious aspects of evaluation practice (Alkin, Vo, & Christie, 2012; Gates, 2017). This literature review explores one aspect of valuing—the valuing involved in reaching a warranted conclusion about the overall merit, worth, or significance of an evaluand. Some theorists refer to this as *synthesis* (e.g., Davidson, 2005; Scriven, 1994).

Scriven (1994) calls for an analytical, reasoned approach to reach a “final synthesis”. He defines synthesis as “the process of combining a set of ratings or performances on several dimensions or components into an overall rating” (Scriven, 1991, p. 342). Stake et al. (1997) calls for a holistic approach to synthesis. He defines synthesis as “the act of putting together, assembling the parts; making sense of complex issues” (Stake, 2004, p. 160). Others question whether an overall conclusion is either necessary or desirable (e.g., Mabry, 1997; Shipman, 2012); or the role of the evaluator in making it (Lincoln & Guba, 2013). The question of whether results should be synthesised into a single evaluative judgement remains unresolved (Shadish, Cook, & Leviton, 1991).

This article discusses the literature on overall evaluative conclusions as an element of evaluation practice. It seeks to inform evaluation practice by answering three questions: 1) How essential is an overall evaluative conclusion, and how much agreement is there

about this aspect of evaluation practice? 2) If an overall conclusion is made, who should take responsibility for it—the evaluator alone or with stakeholders, or stakeholders alone? 3) What methods are available to reach a warranted evaluative conclusion?

Methodology

The main literature search that informed this article was undertaken from June to August 2017. Different combinations of the following search terms were used: *program evaluation*, *valu** (for *valuation* and *valuing*), *judg** (for *judgement*, *judgment*, and *judging*), and *stakeholder*. The broader terms *valuation/valuing* and *judgement/judging* were used instead of the narrower term *synthesis* to ensure that a wider range of literature was assessed for relevance to the topic. The review did not examine published evaluation reports and was not an empirical study of whether evaluation reports include an overall evaluative conclusion. The search focused on programme evaluation, not teacher or student evaluation, product evaluation, or personnel evaluation. The body of literature on cost-effectiveness or cost-benefit analysis was also out of scope of this review.

The Scopus, Discover, and Web of Science databases were searched, as were the databases for specific evaluation journals (*New Directions for Evaluation* and the Sage Journal site that publishes the *American Journal of Evaluation*, *Evaluation Review*, and *Evaluation* journals). Around 80 sources were assessed for their applicability to the review questions. These sources were recorded in a spreadsheet with a brief comment about relevance and key content, and use of a rubric to rate each source: 1 (important and directly relevant to the topic), 2 (relevant), 3 (borderline or somewhat peripheral, but may be included), or 4 (irrelevant with content too distant from the topic) (Student Learning Development University of Otago, 2017). Reference lists were also examined to lead to other relevant sources. In total, 113 sources were

considered, and the final review included over 60 sources.¹ These 60 sources were re-read and notes collated under the broad review questions. Key sources that more substantially or directly addressed each review question were given prominence when synthesising sources to reach a conclusion about each question, but the review also sought to highlight diversity and include less-dominant perspectives.

Davidson (2014a) distinguishes between evaluative reasoning at the micro, meso, and macro levels, where the micro level is evaluation dimensions or criteria, the meso level is synthesising across dimensions to answer evaluation questions, and the macro-level is reaching an overall conclusion about an evaluand. The review started from a focus on a single overall evaluative conclusion at Davidson's (2014a) macro level of evaluative reasoning. It at times widens the scope to valuing at the meso level—that is, synthesising across dimensions to answer evaluative questions, but not necessarily combining these to a single macro-level judgement (Davidson, 2014a). This widening brings more voices to the review, which would not have been included with a strict adherence to a focus on a single synthesis.

Review findings

This article begins by exploring the level of agreement in the literature about the role of an overall evaluative conclusion as an element of evaluation practice. Two positions on valuing generally, and drawing evaluative conclusions more particularly, are contrasted—one that follows criterial logic, and the other that takes a holistic approach. The article then considers who has responsibility for reaching an overall evaluative conclusion—the evaluator or stakeholders? A continuum is used to conceptualise the literature on this topic. The final section of findings considers the literature on the methods available to reach a warranted evaluative conclusion.

¹ This article is based a fuller review, an unpublished paper.

How essential is an overall evaluative conclusion in evaluation practice?

Scriven (1991) and Davidson (2014a) consider that the lack of synthesis is a weakness in evaluation practice. Both see the drawing of an overall conclusion as a key role for the evaluator in nearly all evaluations. Scriven (1991) gives just one example of when a final synthesis may not be appropriate: “when it requires a decision on relative weighting that may be impossible for the evaluator, or will cause unnecessary pain” (p. 343). On these occasions, he suggests that ratings for individual dimensions can be given with an explanation of why the final synthesis has not been undertaken (Scriven, 1994).

Mehrens (1990) suggests that the decisions to be reached after the evaluation determine whether a final synthesis is necessary. His view is that if the evaluation is for the purpose of improvement, an overall judgement is probably not necessary. This differs from Scriven (1994) who suggests that, even in formative evaluation, it is important to report what the evidence adds up to. Davidson concurs with Scriven that “some synthesis is always necessary, whether the evaluation is formative or summative” (Davidson, 2005, p. 153). In a later article she distinguishes between evaluative reasoning at the micro, meso, and macro levels (Davidson, 2014a). All evaluations, irrespective of purpose, require micro-level reasoning to determine the value of each dimension or criteria, and meso-level reasoning to synthesise across multiple dimensions to answer evaluation questions (Davidson, 2014a). Macro-level evaluative reasoning that synthesises answers to all evaluation questions to make an overall conclusion about how worthwhile an intervention is are often, but not always, needed (Davidson, 2014a).

Cronbach (1982) voices concerns about oversimplification that cannot represent the complexity of an evaluand. Stake concurs that “one can squeeze the summary of program quality into a single rating or descriptor, such as A- ... but to ignore the complexity of the

evaluand's activity and merit is to misrepresent the truth and to short-change the audiences" (2004, p. 159). Cronbach (1982) suggests presenting findings at the level of individual findings. Shadish et al. (1991) suggest it might be useful for evaluators to sum up an evaluation with a set of summaries, rather than a single "good/bad" judgement or effect-size estimate (p. 357). Rog (1995) and House (1995) also acknowledge that some evaluators are not comfortable with a "bottom line" judgement and suggest that multiple conclusions or syntheses may be more appropriate. Both consider that a single synthesis can still be aimed for, and that the evaluator can usually arrive at a single value judgement, even if this needs to include qualifying statements (House, 1995; Rog 1995).

In contrast to all these perspectives, postmodernist theorists reject synthesis, resolution, and closure (Abma, 1997; Mabry, 1997). These evaluators seek to identify and express contradictory views of an evaluand and dismiss the trustworthiness of a single set of findings or conclusions (Stufflebeam, 1997). Abma (1997) uses the term "writerly text" to describe an evaluation report that does not include judgements or summaries. Such a text "presents different readings of a situation and allows readers to experience vicariously different points of view. The text is open ... non-linear, the plot recursive, the information partial and incomplete, and interrelations unclear" (Abma, 1997, p. 115). However, postmodernist voices are small in the evaluation literature, which, as the next section describes, is dominated by a rhetoric of criterial logic to reach evaluative conclusions.

The dominant rhetoric: A criterial approach that leads to synthesis

The importance of a warranted argument being developed through a chain of evaluative reasoning that connects the evidence to an evaluative conclusion is emphasised by many theorists (Davidson, 2014a;

Fournier, 1995; Nunns, Peace, & Witten, 2015; Rog, 1995; Schwandt, 2008). Davidson (2014b, p. 38) lists six steps to evaluative reasoning: listing criteria of merit, determining the relative importance of each, defining levels of performance, gathering and analysing evidence, drawing evaluative conclusions about each of the criteria, then finally, synthesising these into overall evaluative conclusions about the programme as a whole. Similarly, much cited is Fournier's (1995) general logic of evaluation which states that all evaluation must first establish criteria of merit, then construct standards for each criterion, then measure performance and compare with the standards, and finally synthesise and integrate data into an overall judgement. Shadish et al. (1991) suggest that much evaluation practice "mimics" the first three steps in Fournier's logic, but less often the final one—synthesis.

Empirical research in Aotearoa New Zealand (Nunns et al., 2015) and Canada (Hurteau, Houle, & Mongiat, 2009) criticises how evaluative conclusions are reached, and the lack of a warranted argument that makes the connection between evidence and evaluative conclusion clear to the reader. Nunns et al. (2015) did a meta-analysis of 30 publicly accessible evaluation reports written or commissioned by public-sector agencies during the period 2010–2013 to understand how evaluative reasoning was being practised in Aotearoa New Zealand. Twenty-four reports contained evaluative conclusions, but only 13 based these on criteria or other comparators defined in the report (Nunns et al., 2015). A similar study was undertaken in Canada, by Hurteau, Houle, and Mongiat (2009). They did a meta-analysis of 40 programme evaluation reports and found that only half went beyond presenting evidence to make a judgement or present a conclusion. None of the reports described how information had been synthesised to make a judgement (Hurteau et al., 2009, p. 313).

The same criticism is made by Davidson (2014a), who describes reading evaluation reports where evaluative conclusions appear to be

based on personal opinion rather than explicit evaluative reasoning. She reviewed half a dozen programme evaluation reports and found that they had made “free associations” to the posed evaluation questions using any data that seemed relevant (Davidson, 2014b, p. 33). Scriven (1991, 1994) also criticises the nature of synthesis in many evaluations, describing it as idiosyncratic, impressionistic, and unreliable:

In many evaluations the lack of any explicit justification of the synthesis procedure is the Achilles heel of the whole effort ... the last step is simply a reaction to the rest, a high-inference judgment which is rather obviously idiosyncratic. (1994, p. 367)

An alternative perspective: A holistic approach

An experiential approach contrasts with the more widespread criterial approach to synthesis (Gates, 2017). This approach is most evident in the work of Stake, particularly his responsive evaluation (Stake, 2004). Stake et al. (1997) reject criterial treatment of an evaluand in favour of a holistic approach that values experiential knowledge. The contrast between this holistic, experiential approach with the more common criterial or analytic approach is most evident in Scriven’s (1994) paper entitled “The Final Synthesis”, presenting the criterial approach, and Stake’s response to it, written with a group of his graduate students (Stake et al., 1997). These papers are cited in other literature as presenting two important perspectives on valuing, and particularly on the drawing of overall evaluative conclusions (Gates, 2017; House & Howe, 1999; Julnes, 2012a, 2012b).

Stake et al. (1997) don’t reject the view that evaluation is about judging the merit or worth of an evaluand, or that this culminates in a synthesis or overall evaluative conclusion. Indeed, they “applaud [Scriven’s] effort to make final syntheses more carefully reasoned” (Stake et al., 1997, p. 90). Where they diverge is in the basis for the synthesis, and the nature of the evaluative reasoning.

Stake (2013) uses metaphor to illustrate the difference between synthesis in responsive evaluation and synthesis in standards-based, criterial evaluation. For the latter, he uses the language of metamorphosis and describes how “the butterfly of ‘value realization’ emerges predictably yet magically from the chrysalis of analysis” (Stake, 2013, p. 179). In his own approach, responsive evaluation, he uses the language of evolution and the metaphor of a crystal to describe how “from the very beginning, shimmers of program value are apparent” (Stake, 2013, p. 179). These metaphors highlight when the synthesis happens—emerging at the end (in the criterial approach), or evolving throughout the evaluation (in the holistic approach).

For Stake, observations and judgements occur simultaneously, and judgements are being made right from an evaluator’s first contact with an evaluand. This experiential approach perceives quality as something that is experienced as well as measured (Stake & Schwandt, 2006). Stake and Schwandt (2006) acknowledge that the relative balance of experiencing and measuring differs depending on context and that the responsive evaluation approach is most suitable when an evaluand can be “intellectually and practically embraced” by an evaluation team (p. 410).

If an overall conclusion is made, who should make it?

The question of who should make an overall conclusion is part of a larger discussion about the role of stakeholders in evaluation. There are three main arguments for including stakeholders in evaluation: a utilisation rationale, an empowerment rationale, and a validity rationale (Mark & Shotland, 1985). The involvement of stakeholders in drawing overall evaluative conclusions could be justified by any of these rationales.

The broader evaluation literature also focuses on the inclusion of stakeholders as a values-based decision reflecting two

dimensions—power and legitimacy (Mark & Shotland, 1985). Evaluators taking a systems-theory approach highlight the need to consider the implications of such decisions rather than take them for granted (Williams, 2014). Who ought to be involved in decision making, and whose expertise is valued or marginalised, are important “boundary choices” that have ethical, political, and practical implications (Williams, 2014).

Those writing about evaluation practice from an indigenous perspective, and evaluators reflecting on evaluation in cultural contexts other than their own, also raise questions of power and privilege when considering the authentic involvement of stakeholders in evaluation (Mertens, 2008; Rogers & Davidson, 2013; Wehipeihana, Davidson, McKegg, & Shanker, 2010). Wehipeihana, a Māori evaluator, highlights the “sense-making” process as being of critical importance when thinking about culture and cultural context in evaluation (Wehipeihana et al., 2010). It is in this sense making that any overall evaluative conclusion is reached. She stresses “the evaluator as a cultural being and the role of personal and cultural values in the drawing of evaluative conclusions and the determination of merit, worth or significance” (Wehipeihana et al., 2010, p. 187). An important question for Wehipeihana is whether the evaluator has the cultural capital to interpret data, both methodologically and ethically.

Davidson (2014a) suggests that all stages in the evaluative reasoning process could be undertaken either collaboratively or by the evaluator only. She poses a number of considerations when deciding who should be involved in evaluative reasoning: *validity* “whose expertise is needed to get the evaluative reasoning right?”; *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)?”; *voice* “who has the right to be at the table when the

evaluative reasoning is done? (indigenous peoples, recipients, community representatives, funders, programme managers)”; and *cost* “at which stages of the evaluative reasoning process is it best to involve different stakeholders for the greatest mutual benefit?” (Davidson, 2014a, p. 8).

Alkin et al. (2012) present a continuum of the possible evaluator and stakeholder roles in valuing. At one end evaluative judgements are the sole responsibility of the evaluator, while at the other the evaluator provides data to stakeholders to make judgements. In between are approaches and evaluation practices where the evaluator guides stakeholders through a valuing process. This continuum is about valuing more broadly, and not just about making an overall evaluative conclusion, but I use it here to frame the discussion about the role of stakeholders in this particular aspect of valuing (see Figure 1).

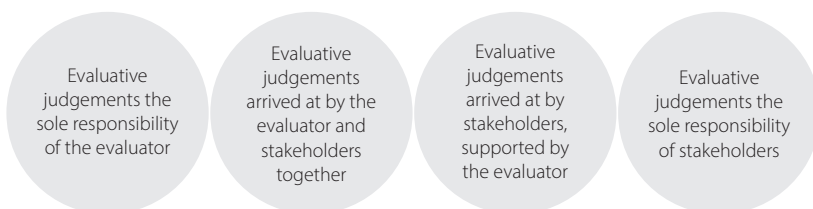


Figure 1. A continuum of the evaluator and stakeholder role in making evaluative judgements, adapted from the typology presented in Alkin [et al.], 2012.

Starting at one end of the continuum, some evaluators view the task of judging merit or worth as the evaluator’s primary and sole responsibility and consider that this should not be passed to stakeholders (Alkin et al., 2012; Christie & Alkin, 2013). Scriven is one such evaluator. In his view, this ensures judgements are unbiased because stakeholders are not involved and can not influence the evaluator’s decision (Alkin et al., 2012). Another perspective on the overall conclusion being the sole domain of the evaluator comes from

methods-based evaluators, where value judgements come directly from the scientific process. These evaluators use experimental designs, such as randomised-controlled trials, to produce a statistically unbiased estimate of the relative effects of a programme, answering questions such as: which programme works best? for whom? and for how long? (Boruch, Solomon, Draine, DeMoya, & Wickrema, 1998) This leads to a conclusion about merit and worth without bias (Alkin et al., 2012; Christie & Alkin, 2013). There may still be a role for the stakeholder in formulating evaluation questions, but the evaluator has responsibility for determining by what methods those questions will be answered (Alkin et al., 2012).

House and Howe's (1999) deliberative democratic approach also focuses on the importance of the evaluator role. While stakeholders are important in the deliberative democratic approach, House and Howe recognise that stakeholder groups do not have equal power, and that dialogue and deliberation will be distorted by power relationships. The evaluator therefore has an important role in using their expertise and the "proper rules and procedures of their discipline" to arrive at valid evaluative conclusions (House & Howe, 1999, p. xvi).

These diverse perspectives illuminate how theorists who see the making of overall evaluative judgements as the role of the evaluator alone do so for different reasons.

A second position, closer to the middle of the Alkin et al. (2012) continuum, maintains a strong role for the evaluator, but the evaluator participates with stakeholders in the valuing process.

Evaluation-specific methodologies—such as rubrics—have come to be widely used in criterial-based evaluations that involve the evaluator and stakeholders working together. Evaluators have an important role in leading the development of rubrics as part of a clear methodology that ensures evaluative conclusions can be reached. However, stakeholder participation is highly valued, and stakeholders have a

high level of involvement in, and influence on, the evaluation, particularly in defining what quality look likes in their particular context (Rogers & Davidson, 2013). A body of practice-informed literature about the use of rubrics is emerging. This discusses what value there is in involving stakeholders in all stages of evaluative reasoning, including making evaluative judgements against individual criteria and synthesising to reach an overall judgement (Davidson, 2014a; King, McKegg, Oakden, & Wehipeihana, 2013; Oakden, 2013). Using rubrics with clients and stakeholders increases the transparency of how such judgements have been reached (King et al., 2013).

A third position on the continuum is evaluators who see the making of overall judgements or conclusions as the domain of stakeholders, but still describe a role for the evaluator in facilitating or guiding this. Alkin likes to work with stakeholders to support them to draw evaluative conclusions by presenting simulated outcomes or hypothetical findings at the outset of an evaluation, before data have been collected (Alkin et al., 2012). He does not believe the evaluator should impose their own value judgements on data, and prefers to present evaluation data “factually” without making judgements about value (Alkin, 2013; Alkin et al., 2012; Christie & Alkin, 2013). This allows the valuing to be done by stakeholders. Guba and Lincoln also give stakeholders primacy in assigning value, and describe the role of the evaluator as being to “facilitate negotiations between individuals reflecting multiple realities” (Christie & Alkin, 2013, p. 39). In this constructivist approach to evaluation, stakeholders have access to all data, and findings are arrived at by joint construction and negotiation by stakeholders, with the evaluator as facilitator (Lincoln & Guba, 2013).

Finally, at the furthest end of the continuum are theorists who advocate that making evaluative judgements is not the role of the evaluator at all. The basis for this view is that the value of something

changes depending upon the context, including the values and perspectives of individuals (Alkin et al., 2012). Alkin et al. propose two types of scenario. In the first the evaluator provides the data for known stakeholders to draw their own conclusions without facilitation. In the second the valuing or drawing of conclusions is left to “general/broad others”, such as the reader of an evaluation report. So, in essence, this means that an overall evaluative judgement is not made.

What methods are available to reach a warranted evaluative conclusion?

This final section considers the evaluation literature on methods for making an overall evaluative conclusion. Alkin et al. (2012) conclude that literature that operationalises ways to reach these judgements is sparse, although Davidson (2005) suggests there has been good progress made in evaluation-specific methodologies for synthesising. Davidson herself has been influential in building evaluation capability in these methodologies in Aotearoa New Zealand and in North America (King et al., 2013).

Julnes (2012b) gives an overview of four different methods for synthesis. The first—minimal aggregation—is a method that involves not doing synthesis or making an overall evaluative judgement but reporting performance on multiple criteria separately. The second—a checklist approach—is described as: identifying criteria, establishing standards, displaying performance for each criterion, and then “reaching qualitative or quantitative syntheses of ranking and ratings accordingly” (Julnes, 2012b, p. 9). This method follows the criterial logic discussed earlier in this literature review, but Julnes gives no detail on how to use this synthesis method in practice. The third aggregation method Julnes describes is quantitative aggregation—weighting performance on each criterion, then doing

a mathematical calculation to get an overall score. The fourth and final method he describes is social aggregation, which is “a stark contrast to quantitative synthesis” (Julnes, 2012b, p. 10). House and Stake are cited as two evaluators that take this approach to synthesis, using deliberation and “disciplined intuition” respectively to support judgements (Julnes, 2012b, p. 10). House’s five stages in the deliberative process align with the four steps in Fournier’s general logic of evaluation (1995), although House adds “intuitive to-and-fro reasoning” (1995, p. 39).

Another distinction between synthesis methods is whether synthesis is statistical or clinical (Mehrens, 1990). Statistical synthesis combines data mathematically (Mehrens, 1990). In contrast, taking a clinical or judgemental approach is described as “eyeballing the various pieces of data to arrive at a final score or decision” (Mehrens, 1990, p. 324). Julnes (2012a) makes a similar distinction but uses the terms *algorithmic* and *holistic aggregation*, one being “adding up the judgements on a checklist and reporting sums” and the other “a more informal, intuitive approach” (p. 120).

Scriven’s writings (1991, 1993, 1994, 1995, 2012) provide a foundation for understanding synthesis as the process of combining performance on multiple dimensions into an overall rating. His logic culminates in the evaluator making highly probable probative inferences—“inference to a conclusion that has been established, so the utterer claims and is prepared to support, as beyond reasonable doubt” (Scriven, 2012, p. 23).

Davidson (2005) builds on Scriven’s contribution to developing an evaluation-specific logic and methodology. In her book on evaluation methodology, Davidson (2005) devotes a chapter to synthesis methodology, describing different types of synthesis methodology in considerable detail and working through examples. This includes qualitative and quantitative methodology for two synthesis purposes:

synthesis for ranking (e.g., comparing three pilot programmes) and synthesis for grading (e.g., deciding if a single programme was worth implementing).

The quantitative weight-and-sum method requires evaluators to numerically combine importance and performance on each dimension of interest, then multiply weights by performance scores before summing these products. This gives the overall merit of the evaluand (Davidson, 2005, p. 172). However, numerical weighting systems “often lead to conclusions that have the evaluation team staring at a conclusion that seems not to be quite right, thereby leading to the temptation of juggling the weights until the answer looks right.” (Davidson, 2005, p. 157). Shadish et al. (1991) suggest that a particular problem with weight-and-sum methods is when findings may conflict—for example, when a programme has benefits for some groups but negative consequences for others. Davidson (2005) suggests that qualitative synthesis methodology can be more appropriate. This is where qualitative labels are applied, and performance on each dimension is combined without the use of multiplication. In later publications, Davidson (2014a, 2014b) develops this methodology further and represents it visually with a diagram of a funnel. The funnel is evaluative interpretation that evidence goes into. Inside the funnel, the evidence is looked at evaluatively, and evaluative conclusions come out (Davidson, 2014b).

Rubrics can be used as part of a synthesis methodology (King et al., 2013; Oakden, 2013; Oakden & Weenink, 2015; Stone-Jovicich, 2015). Rubrics comprise evaluative criteria, which are the aspects of performance the evaluation focuses on, and the description of what performance of the criterion looks like at different levels (Oakden, 2013). Rubrics sometimes use the concept of hurdles or bars in making an overall evaluative conclusion (Davidson, 2005; Scriven, 1991). A bar is a minimum level of performance on a dimension where low

performance cannot be compensated for by better performance on other dimensions (Davidson, 2005, p. 166). A *hard hurdle* (Davidson, 2005) or *global bar* (Scriven, 1991) is an overall pass requirement for the evaluand, and a *soft hurdle* is where a poor rating on a dimension puts a limit on the overall rating that can be achieved (Davidson, 2005). These can also be framed as “deal breakers” (Oakden, 2016).

Eoyang and Oakden (2016) view evaluation-specific methodology through a human systems dynamics lens. One strategy used in the synthesis stage of their adaptive evaluation approach is “pattern spotting”, which originates from the work of Capper and Williams (2004). After analysis has been done using a generic rubric at the level of each criterion, they work through five stages, often with stakeholders. The first takes a broad overview of the data overall to identify key generalisations. It asks: “in general, what is this data telling us?” (Eoyang & Oakden, 2016, p. 11). The second stage looks for exceptions to each generalisation, such as outliers—really excellent or poor ratings—that need to be taken into account. The third and fourth stages look for contradictions and things that are surprising in the data. The final stage considers what is still puzzling and seeks to explore those puzzles rather than explain them away (Eoyang & Oakden, 2016).

Stake et al. (1997) also highlight the importance of looking for patterns. They advocate for “disciplined intuition”, describing intuition as “pulling hidden patterns together so that we can make better judgments about the evaluand” (Stake et al., 1997, p. 96). Judgement is necessarily subjective rather than mathematical:

In the final synthesis, there is need for both intuition and conscious reasoning ... [S]hifting back and forth between the formal and the informal, the general and the particular, the hunch and the habit, increases the strength and comprehensiveness of this scrutiny. (Stake et al., 1997, p. 92)

Implications for evaluation

This review has concluded that there are diverse perspectives about how an overall evaluative conclusion should be made, on what basis, who should be involved, and what methods should be used. There is no consensus on the overall evaluative conclusion as an aspect of evaluation practice. This is not surprising, given that “there are multiple, defensible options for attending to each issue in evaluation practice” (Gates, 2017, p. 8). All evaluations have their own unique blend of theory and practice, and require trade-offs between evaluation purpose, stakeholder needs, degree of certainty required in conclusions, budget, and time available (Davidson, 2000; Shadish, 1998). Alkin et al. conclude that the evaluator engages in a process, a major element of which is assuring that valuing will occur. We must recognize the diversity of contexts, both evaluation context and evaluator context, and not insist that all evaluators perform valuing in one (presumed) ‘right’ way. (2012, p. 39)

Others (e.g., Cronbach, 1982; Julnes, 2012a; Patton, 2012) also emphasise that context matters, and that decisions about how to go about valuing in an evaluation are context dependent. So what does that mean for evaluation practice? Does anything go?

This review sought to add to my own—and others’—inventory of conceptual knowledge about this area of theory and practice (Schwandt, 2014). A fuller inventory gives evaluators more “to think with”, and a richer “evaluation imagination” to draw on, when making decisions about evaluation practice (Schwandt, 2014, p. 236). Knowledge of diverse perspectives and theoretical positions gives evaluators options and alternatives when considering the place for an overall evaluative conclusion or synthesis in their next evaluation. This supports and enhances reflective practice, a domain in the Aotearoa New Zealand Evaluation Association (2011) competencies.

There are two important implications from these findings—the first for research on evaluation, given the limited empirical research on the topic, and the second for evaluator professional development. Julnes (2012a) calls for more empirical evidence to understand the impact of different valuing methodologies. Nunns et al. (2015) and Hurteau et al. (2009) have sought to understand evaluation practice around warranted evaluative conclusions in Aotearoa New Zealand and Canada respectively. Each reports a snapshot of practice, which, although useful, does not address Julnes’ call to understand the implications for evaluation use or validity. Absent from the literature is research that explores client or stakeholder views on the inclusion or exclusion of an overall conclusion or synthesis, and how this influences their engagement with and use of evaluation findings. Gates (2017) concludes that the making of value judgements is still one of the least clear-cut aspects of evaluation practice.

Patton (2012) asks us to examine our own evaluator context and poses three questions relevant to this review topic: “What influences the way in which you go about the valuing process? What are your theoretic dispositions as an evaluator? How does your individual point of view and preferences inform and undergird your approach to the valuing process?” (p. 105). Evaluator competencies support and anchor such reflective practice by providing a framework for self-assessment (Stevahn, King, Ghere, & Minnema, 2005, p. 46).

Synthesis is explicitly mentioned within two domains of the Aotearoa New Zealand competencies. Domain 1, contextual analysis and engagement, suggests evaluators need to demonstrate “ability to bring the contextual analysis and engagement together so that the evidence, analysis, synthesis and evaluative interpretation of the information gathered is credible and valid to the range of people (stakeholders) involved in and affected by the evaluation” (Aotearoa New Zealand Evaluation Association, 2011, p. 13). Domain 2, systematic

evaluative inquiry, refers to synthesis in four places, including evaluators having knowledge, skill, and ability to systematically synthesise information and interpret findings to reach “valid, defensible, and transparent conclusions and/or judgements and answers to evaluative questions” (Aotearoa New Zealand Evaluation Association, 2011, p. 13). The competencies also highlight that evaluators bring critical thinking, analytical, and synthesis skills to evaluation (Aotearoa New Zealand Evaluation Association, 2011).

There is a call for competencies about valuing generally, and synthesis specifically, to be given attention in evaluator training (Gullickson, 2017; Julnes, 2012a). Davidson (2014b) identifies that synthesis methodology is not taught in social science research training and Gullickson concludes that there is little-to-no emphasis on synthesis in evaluator teaching or training. Given that this review shows that when synthesis is done it is not always done well (Davidson, 2014b; Hurteau et al., 2009; Nunns et al., 2015), and that literature on different ways to reach an overall evaluative conclusion is sparse, this review adds to the call for this area of evaluation practice to be highlighted in evaluator training and professional development. This could include how to deliberate over findings and conclusions when diverse stakeholders are actively involved (House & Howe, 1999; McDonald, 2008). McDonald considers an evaluator needs “courage to make clear judgements, backed by solid evidence, with respect to the evaluand, even though the conclusions drawn might not be to the liking of some stakeholders ... [and] the capacity to facilitate ‘difficult’ decisions between stakeholder groups about the findings” (2008, p. 264).

Conclusion

This literature review has attempted to represent the breadth of thinking across the evaluation literature about making an overall evaluative conclusion. Two prominent evaluation theorists, Scriven

and Stake, are central to considering this aspect of evaluation practice. This review sought to also bring other perspectives to the discussion. The review concludes that there has been little empirical research undertaken on the valuing involved in reaching a warranted conclusion about the overall merit, worth, or significance of an evaluand. The literature does not definitively inform evaluators as to whether they should be making an overall evaluative conclusion, who they should involve, and what methods they should use. As evaluators, we “generally accept that we have multiple methods in our collective toolkits, and that methods must fit their contexts; this applies equally to methods of valuing” (Julnes, 2012a, p. 125). This review supports evaluators to increase their theoretical knowledge, to reflect on their practice, and to consider the diversity of perspectives about the valuing involved in an overall evaluative conclusion. It encourages evaluators to make deliberate, informed decisions about the making—or not—of an overall conclusion in their future evaluation practice.

References

- Abma, T. A. (1997). Sharing power, facing ambiguity. In L. Mabry (Ed.), *Evaluation and the postmodern dilemma* (Vol. 3, pp. 105–119). Greenwich, CT: JAI Press.
- Alkin, M. C. (2013). Context-sensitive evaluation. In M. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed., pp. 283–292). Thousand Oaks, CA: Sage.
- Alkin, M. C., Vo, A. T., & Christie, C. A. (2012). The evaluator's role in valuing: Who and with whom. *New Directions for Evaluation*, 133, 29–41. <https://doi.org/10.1002/ev.20004>
- Aotearoa New Zealand Evaluation Association. (2011). *Evaluator competencies*. Retrieved from https://www.anzea.org.nz/app/uploads/2019/04/110801_anzea_evaluator_competencies_final.pdf

- Boruch, R., Solomon, P., Draine, J., DeMoya, D., & Wickrema, R. (1998). Design-based evaluations: Process studies, experiments, and quasi-experiments. *International Journal of Social Welfare*, 7(2), 126. <https://doi.org/10.1111/j.1468-2397.1998.tb00213.x>
- Capper, P., & Williams, B. (2004). *Enhancing evaluation using systems concepts CHAT*. Paper presented at the American Evaluation Association Conference, Atlanta, GA. Retrieved from <http://www.bobwilliams.co.nz/ewExternalFiles/activity.pdf>
- Christie, C., & Alkin, M. C. (2013). An evaluation theory tree. In M. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed., pp. 11–57). Thousand Oaks, CA: Sage.
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Josey-Bass.
- Davidson, E. J. (2000). Ascertaining causality in theory-based evaluation. *New Directions for Evaluation*, 87, 17–26. doi: 10.1002/ev.1178
- Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage. <https://doi.org/10.4135/9781452230115>
- Davidson, E. J. (2014a). *Evaluative reasoning. Methodological briefs: Impact evaluation 4*. Florence, Italy: UNICEF Office of Research. Retrieved from https://www.unicef-irc.org/KM/IE/impact_4.php
- Davidson, E. J. (2014b). How “beauty” can bring truth and justice to life. *New Directions for Evaluation*, 142, 31–43. <https://doi.org/10.1002/ev.20083>
- Eoyang, G., & Oakden, J. (2016). Adaptive evaluation. *Emergence: Complexity & Organization*, 18(3/4), 1–14.
- Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. *New Directions for Evaluation*, 68, 15–32. <https://doi.org/10.1002/ev.1017>

- Gates, E. F. (2017). Toward valuing with critical systems heuristics. *American Journal of Evaluation*, Prepublished 21 July 2017. <https://doi.org/10.1177/1098214017703703>
- Gullickson, A. (2017, November). *Evaluative synthesis as a tool for teaching and assessing evaluative thinking*. Paper presented at the American Evaluation Association Conference, Washington, DC.
- House, E. (1995). Putting things together coherently: Logic and justice. *New Directions for Evaluation*, 68, 33–48. <https://doi.org/10.1002/ev.1018>
- House, E., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.
- Hurteau, M., Houle, S., & Mongiat, S. (2009). How legitimate and justified are judgments in program evaluation? *Evaluation*, 15(3), 307–319. <https://doi.org/10.1177/1356389009105883>
- Julnes, G. (2012a). Developing policies to support valuing in the public interest. *New Directions for Evaluation*, 133, 109–129. <https://doi.org/10.1002/ev.20012>
- Julnes, G. (2012b). Managing valuation. *New Directions for Evaluation*, 133, 3–15. <https://doi.org/10.1002/ev.20002>
- King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Rubrics: A method for surfacing values and improving the credibility of evaluation. *Journal of MultiDisciplinary Evaluation*, 9(21), 11–20.
- Lincoln, Y. S., & Guba, E. G. (2013). The roots of fourth generation evaluation. In M. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed., pp. 218–228). Thousand Oaks, CA: Sage.
- Mabry, L. (1997). A postmodern text on postmodernism. In L. Mabry (Ed.), *Evaluation and the postmodern dilemma* (Vol. 3, pp. 1–19). Greenwich, CT: JAI Press.
- Mark, M. M., & Shotland, R. L. (1985). Stakeholder-based evaluation and value judgments. *Evaluation Review*, 9(5), 605–626. <https://doi.org/10.1177/0193841X8500900504>

- McDonald, D. (2008). Revisiting a theory of negotiation: The utility of Markiewicz (2005) proposed six principles. *Evaluation and Program Planning*, 31(3), 259–265. <https://doi.org/10.1016/j.evalprogplan.2008.03.003>
- Mehrens, W. A. (1990). Combining evaluation data from multiple sources. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation* (pp. 322–334). London, UK: Sage.
- Mertens, D. M. (2008). Stakeholder representation in culturally complex communities: Insights from the transformative paradigm. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 41–60). New York, NY: The Guilford Press.
- Nunns, H., Peace, R., & Witten, K. (2015). Evaluative reasoning in public-sector evaluation in Aotearoa New Zealand: How are we doing? *Evaluation Matters—He Take To Te Aromatawai*, 1, 137–163. <https://doi.org/10.18296/em.0007>
- Oakden, J. (2013). Evaluation rubrics: How to ensure transparent and clear assessment that respects diverse lines of evidence. *BetterEvaluation*. Retrieved from <http://betterevaluation.org/resource/example/rubrics-oakden>
- Oakden, J. (2016, September). *Evaluating in complexity: What have we learned about designing and running evaluations?* Paper presented at the American Evaluation Association Conference, Atlanta, GA. Retrieved from <http://pragmatica.nz/wp-content/uploads/pragmatica-nz/sites/326/161028-Oakden-AEA-jo-xx.pdf>
- Oakden, J., & Weenink, M. (2015, July). *What's on the rubric horizon: Taking stock of our current practice and thinking about what is next.* Paper presented at the Aotearoa New Zealand Evaluation Association conference, Auckland. Retrieved from <http://pragmatica.nz/wp-content/uploads/pragmatica-nz/sites/326/150718-Oakden-Weenink-ANZEA-vxx.pdf>
- Patton, M. (2012). Contextual pragmatics of valuing. *New Directions for Evaluation*, 133, 97–108. <https://doi.org/10.1002/ev.20011>

- Rog, D. J. (1995). Reasoning in evaluation: Challenges for the practitioner. *New Directions for Evaluation*, 68, 93–100. <https://doi.org/10.1002/ev.1022>
- Rogers, P., & Davidson, E. J. (2013). Australian and New Zealand evaluation theorists. In M. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed., pp. 371–385). Thousand Oaks, CA: Sage.
- Schwandt, T. A. (2008). Educating for intelligent belief in evaluation. *American Journal of Evaluation*, 29(2), 139–150. <https://doi.org/10.1177/1098214008316889>
- Schwandt, T. A. (2014). On the mutually informing relationship between practice and theory in evaluation. *American Journal of Evaluation*, 35(2), 231–236. <https://doi.org/10.1177/1098214013503703>
- Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.
- Scriven, M. (1993). Intermediate evaluation design issues. *New Directions for Program Evaluation*, 58, 67–74. <https://doi.org/10.1002/ev.1643>
- Scriven, M. (1994). The final synthesis. *Evaluation Practice*, 15(3), 367–382. <https://doi.org/10.1177/109821409401500317>
- Scriven, M. (1995). The logic of evaluation and evaluation practice. *New Directions for Evaluation*, 68, 49–70. <https://doi.org/doi:10.1002/ev.1019>
- Scriven, M. (2012). The logic of valuing. *New Directions for Evaluation*, 133, 17–28. <https://doi.org/10.1002/ev.20003>
- Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation*, 19(1), 1-19. doi: 10.1177/109821409801900102
- Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation*. Newbury Park, CA: Sage.
- Shipman, S. (2012). The role of context in valuing federal programs. *New Directions for Evaluation*, 133, 53–63. <https://doi.org/10.1002/ev.20006>
- Stake, R. (2004). *Standards-based & responsive evaluation*. Thousand Oaks, CA: Sage. <https://doi.org/10.4135/9781412985932>

- Stake, R. (2013). Responsive evaluation IV. In M. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (2nd ed., pp. 189–197). Thousand Oaks, CA: Sage.
- Stake, R., Migotsky, C., Davis, R., Cisneros, E. J., Depaul, G., Dunbar, C., ... Chaves, I. (1997). The evolving syntheses of program value. *Evaluation Practice*, 18(2), 89–103. <https://doi.org/10.1177/109821409701800202>
- Stake, R., & Schwandt, T. (2006). On discerning quality in evaluation. In I. Shaw, J. C. Greene, & M. Mark (Eds.), *The SAGE handbook of evaluation* (pp. 404–418). London, UK: Sage.
- Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation*, 26(1), 43–59. <https://doi.org/10.1177/1098214004273180>
- Stone-Jovicich, S. (2015) To rubrics or not to rubrics? *Practice Notes Series: Managing for Impact*. Dickson, ACT: The Food Systems Innovation.
- Student Learning Development University of Otago. (2017). *Writing a literature review*. Retrieved from <https://www.otago.ac.nz/hedc/otago615355.pdf>
- Stufflebeam, D. L. (1997). A standards-based perspective on evaluation. In L. Mabry (Ed.), *Advances in program evaluation: Evaluation and the postmodern dilemma* (Vol. 3, pp. 61–88). Greenwich, CT: JAI Press.
- Wehipeihana, N., Davidson, E. J., McKegg, K., & Shanker, V. (2010). What does it take to do evaluation in communities and cultural contexts other than our own? *Journal of MultiDisciplinary Evaluation*, 6(13), 182–192. Retrieved from http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/265/256
- Williams, B. (2014). Prosaic or profound? The adoption of systems ideas by impact evaluation. *IDS Bulletin, Special Issue: Towards Systemic Approaches to Evaluation and Impact*, 46(1), 7–16. <https://doi.org/10.1111/1759-5436.12117>

The Author

Jo MacDonald is a senior researcher at the New Zealand Council for Educational Research (NZCER).

Email: jo.macdonald@nzcer.org.nz