Journal of MultiDisciplinary Evaluation
Volume 18, Issue 42, 2022

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Quality as Praxis: A Tool for Formative Metaevaluation

Amy C. Jersild
Western Michigan University

Michael A. Harnar
Western Michigan University

Background: While evaluation theorists speak to the importance of formative metaevaluation, examples of how to do this are rarely specified in the evaluation literature. How evaluators engage in formative metaevaluation practice is not fully known or researched.

Purpose: This paper aims to (a) further explore formative metaevaluation as a means for quality assurance, with implications for both developing evaluators' capacity and advancing evaluation as a field of practice; and (b) present a tool with the intent to move toward a more deliberate formative quality evaluation practice.

Setting: Development of a baseline, formative and summative evaluation for a human trafficking program in South Asia.

Intervention: NA

Research Design: Auto-ethnographic approach

Data Collection and Analysis: NA

Findings: Discussion focuses on the relationship between evaluator and commissioner and how the development and use of a deliberate approach to formative metaevaluation, through examination of a proposed tool, can lead to a clearer definition of evaluation quality. Formative metaevaluation can be an important tool for evaluators in exercising professional judgment and in taking an active role in advancing the evaluation field.

Keywords: *meta-evaluation; metaevaluation; meta evaluation; quality; quality assurance; formative meta-evaluation.*

Introduction

Michael Scriven (1991) describes metaevaluation as important to the field of evaluation from both a scientific and a moral perspective. Particularly where others are impacted by the evaluation process□which may be the case for most evaluations□he notes metaevaluation as an ethical imperative and indeed a "professional imperative" (p. 229). Asking "Who evaluates the evaluator?" reminds us that evaluation is a self-referential activity. Thomas Schwandt (2015) describes evaluators' commitment to metaevaluation in similarly ethical terms, as an endeavor in "practicing what we preach" (p. 134).

Other evaluation theorists have also highlighted the importance of metaevaluation research and practice since Scriven (1969) coined the term (e.g., Henry & Mark, 2003; Fitzpatrick et al., 2011; Stufflebeam & Shinkfield, 2007; Stufflebeam & Coryn, 2014). Dahler-Larsen (2019) reminds us that if we do not continuously reflect on and assess the level of quality in our work, the original purpose and intent of the system or field in which we operate may slowly be lost. Metaevaluation is a means for the evaluation field to identify and ensure quality in practice, and, as Dahler-Larsen points out, how we research and practice quality evaluation matters.

Jacobs and Affrodegon (2015) note that the evaluation field has experienced multiple generations of metaevaluation. They identify our current as a fourth generation or "maturity period," which coincides with greater levels of practice and the operationalization of instruments for use. These instruments include standards (e.g., Yarbrough et al, 2010; SEVAL, 2016; SCE, 2011) and directing principles proliferated by organizations (e.g., AEA guiding principles; UNEG guidelines).

At an operational level, Stufflebeam and Coryn (2014) elaborate on a practical definition of metaevaluation as either formative or summative and either external or internal. Proactive (or formative) metaevaluations help evaluators focus, design, budget, contract, and carry out sound evaluations, and retrospective metaevaluations (or summative metaevaluations) help audiences judge completed evaluations. Each may be carried out either internally or externally, and the authors stress the importance of both the evaluator and the commissioner being engaged. Despite explicit calls for more metaevaluation research and practice (e.g., Henry & Mark, 2003; Stufflebeam & Coryn, 2014), it is understood that summative metaevaluation practice is far more prevalent and more often discussed in the evaluation literature than is formative metaevaluation practice. Further, summative evaluation practice is often performed externally and not by evaluators themselves (Cooksy & Caracelli, 2009).

In responding to the direct call for more metaevaluation, and to the dearth of research on how formative metaevaluation is or should be practiced, we have, using an auto-ethnographic method, devised a tool that offers an additional and alternative approach to formative metaevaluation practice. We have firmly rooted it in quality as practice in a pre-formative sense. By pre-formative, we note Scriven's (2012) reference to work that is done with "the purpose to improve the merit, worth, or significance of a possible evaluand" (p. 59). We aim to (a) further explore formative metaevaluation as a means for quality assurance, with implications for both developing evaluators' capacity and advancing evaluation as a field of practice; and (b) in response to the evaluation literature's explicit emphasis on the importance of formative metaevaluation to the profession, to present a tool that, with further testing and refinement, may move the evaluation field toward a more deliberate formative quality evaluation practice. By addressing these objectives, we explore the concept of building out quality assurance in evaluation, as well as how and to what standards quality is measured.

Quality as Practice

The notion of practice as quality can be found in the writings of both Schwandt and Dahler-Larsen. Schwandt (2003) speaks to practice and the practical in evaluation as both a bilateral and an interactive pursuit, one that involves both self-knowledge as an evaluator and awareness of how one is known and perceived by others. Practice, according to Schwandt, involves "perception and practical reason," a regard for "warrants, values, emotions and commitment" (p. 355), and that which is "indispensable to evaluative judgment" (p. 356). This particular kind of knowledge or practical reason goes beyond the technical knowledge evaluators have about research and evaluation methods and instead aims to answer value-rational questions, such as "How should I be in this situation? What should be done? Is this desirable?" (p. 354).

It is an iterative praxis that evolves as one goes. It demands a particular kind of practical reason, Schwandt (2003) notes, and a knowledge that is self-aware. The end goal is not control over an object or product but the knowing of how to "function together with those with whom I am engaged in understanding, deciding, and acting" (p. 356).

Quality as practice is one of the nine perspectives of quality Dahler-Larsen (2019) describes. Dahler-Larsen points to Schwandt's application of Aristotle's notion of phronesis (practical wisdom or judgment) and Dewey's notion of situated qualitative thinking. Experience and knowledge of what is typical in an evaluation context, along with the ability to respond, are the marks of a good practitioner. Interestingly, Dahler-Larsen frames this kind of knowledge as akin to the craftsmanship of a learned musician, artist, or athlete. Thus, he distinguishes quality as practice from another of the nine types he discusses: quality as compliance with a particular set of criteria and standards. He further notes that quality as practice may not be easily planned or implemented; rather, it is intuitive and improvisational, a characterization that resonates with Schwandt's notions of perception and practical reasoning.

To build on the craftsmanship metaphor Dahler-Larsen alludes to, one may also look to the writings of the renowned Japanese craftsman and ceramicist, Soetsu Yanagi, who wrote about quality as beauty (Yanagi & Leach, 1972). Yanagi describes the quiet familiarity of handmade craft objects, their feel in one's hand, and their facilitation of use and function,. These qualities may be thought of as akin to evaluation's practice and process-oriented collaboration; its orientation around an outcome of that process (whether via a report or some other form of expression); and its subsequent use and engagement. Yanagi's notion of quality as beauty, like the idea of evaluation as craft, rests upon perception and appreciation, as well as a balanced interconnectedness.

Quality Assurance in Evaluation

How, practically, do evaluators build quality assurance into their practice? Along with extensive other guidance, the evaluation literature offers metaevaluation (generally by way of evaluation standards, principles, and checklists) as a means to ensure quality (Fitzpatrick et al., 2011; Patton, 2008; Scriven, 2009; Stufflebeam, 1999, 2001a, 2001b; Sanders, 1995).

Scriven (1991) notes that checklists "provide an extremely versatile instrument for determining the quality of many kinds of work, programs, activities, and products, and may be used to guide observation or a series of measurement efforts" (p. 80). Stufflebeam (2001b) describes an evaluation checklist as "a list for guiding an enterprise to success (formative orientation) and/or judging its merit and worth (summative orientation)" (p. 171). Most checklists present a framework that, either explicitly or implicitly, makes a claim to comprehensiveness (Scriven, 2009). A well-crafted checklist breaks down concepts to facilitate reliable recall, understanding, and assessment. Yet some have speculated about whether evaluators' tendency to use checklists in a ritualistic fashion may reduce checklists' reliability (Scriven, 2009).

The Joint Committee on Standards for Educational Evaluation's Program Evaluation Standards (Yarbrough et al, 2010) provide the evaluation field with a set of aspirational standards in five areas: feasibility, propriety, utility, accuracy, and accountability. The standards form the systematic basis that Scriven (2009) describes for the evaluation checklists noted above. The Joint Committee (Yarbrough et al, 2010) prescribes the application of these standards in practice as an intuitive process that requires deep understanding of the standards, calling to mind the knowledge, practical reason, and self-awareness Schwandt (2003), as discussed above, describes.

Harnar et al. (2020) found among a sample of American Evaluation Association members that evaluators are concerned with the credibility and reliability of their work and will compare components of their work against a variety of standards to strengthen their warrants about their evaluations' credibility and validity. Many will engage stakeholders throughout the life cycle of an evaluation to ensure expectations are being met. Evaluators in this study described a relatively subjective and intuitive process for judging adherence to standards. Much of what respondents described could be interpreted as procedures "for describing an evaluation activity and judging it against a set of ideas concerning what constitutes good evaluation" (Stufflebeam, 2011, p. 135), or an informal formative metaevaluation.

The evaluation literature's focus on metaevaluation as an important means for maintaining and ensuring quality in evaluation practice (and for professionalizing the evaluation field) raises a question of interest: Whose standards—or whose interpretation of standards—are to prevail in how we as an evaluation field define quality? The Program Evaluation Standards' reliability and validity as a decision-making tool cannot be taken for granted. Wingate (2009), analyzing a group of evaluators' use of the evaluation standards to examine evaluation reports, found low interrater reliability; evaluators' interpretations of the standards reflected their varied knowledge and experience.

The varying interpretation of standards found by Wingate (2009) and the relatively subjective adherence to standards found by Harnar et al. (2020) are further complicated by an environment of economic exchange, where an evaluator (supplier) provides evaluation services to an organization or agency (client). The concepts of intrinsic quality (or quality as defined by the evaluator), and extrinsic quality (or quality as defined by the client) (Fitzpatrick et al., 2011; Harnar et al., 2020) can frame several questions: Whose set of standards, or interpretation thereof, are we to be mindful of? Whose priorities matter during the evaluation, and what are they? Evaluation as a marketplace activity, a service that is commissioned for funding, can often leave determination of quality to the commissioner, the party who pays for the service. Indeed, as noted above, summative metaevaluation is much more commonplace than formative, and is most often done by the commissioner or contracted to a third party (e.g., UNICEF's global evaluation reports oversight system (GEROS), https://www.unicef.org/evaluation/global-evaluation-reports-oversight-system-geros).

To ensure that evaluation continues to grow as a profession, and to ensure that evaluators are not merely technicians responding to demands, evaluators can and need to lead dialogue on quality practice, rooted within standards and principles, both among ourselves and with commissioners. This will contribute toward moving evaluation from what Picciotto (2011), quoting Lincoln (1985), calls a fledging profession to a more solid and recognized one supported by autonomous professional guidelines, ethical standards, and quality assurance. Fee dependency remains a challenge to evaluator independence and therefore an issue for the evaluation field (Picciotto, 2011, 2020), and assuming control over our professional standards and ethical guidelines is imperative to professional standing and growth. As Picciotto (2011) notes, "The status of any expert occupation is best understood in terms of the sources of power and authority over the definition and control over specialized knowledge work" (p. 174). Indeed, we would argue that in moving from a market or managed good to a more established knowledge profession (as discussed in the sociology of professions literature, e.g., Abbott, 1988; Friedson, 1983), the evaluation profession will only become a full-fledged profession when it has control over its work as well as the standards by which the work is evaluated. Thus it behooves us to better understand and research quality and formative metaevaluation practice, and to take ownership of definitions of quality evaluation practice.

The questions remain: How do we adopt the prescriptive guidance in the literature on metaevaluation? How do we use standards in a balanced way as evaluators mindful of quality practice? How do we apply the practical wisdom and knowledge that Schwandt and Dahler-Larsen so eloquently write about in a purposeful way? We set out to answer these questions by developing a tool for use while engaged in a large evaluation project. The axiological assumptions (discussed above) of evaluator independence and advancement of the evaluation field underlie our pre-formative metaevaluation approach. Our intent is to make explicit our consideration of the standards and principles for quality evaluation, and strive for what Yarbrough et al. (2010) identify as "accountable evaluation," a process of "documenting how specific standards have been selected and implemented and which trade-offs were required to balance effectiveness and efficiency" (p. xxxviii). Herein we describe our tool and conclude with reflection on its practical use and value.

Context for Tool Development

In 2019 a bilateral donor funded a foundation to support work on modern slavery programming in two neighboring countries in Asia to reduce human trafficking prevalence through new and innovative approaches and scaling of tested approaches. The foundation issued a request for proposals and funded 10 projects in the region to implement a range of interventions focused on different types of services and sectors. The donor funded a third organization specialized in monitoring and evaluation to provide a set of four interlinked services to the foundation (monitoring, adaptive management, evaluation, and learning) over a three-year period. This unit was known as the monitoring, evaluation and learning unit (MELU). Amy C. Jersild, the lead author of this paper, was contracted to be the lead evaluator.

The evaluation team launched a five-month inception phase to plan a baseline study, formative and summative evaluations. We held a series of workshops and consultations, analyzed available program documents, and reviewed the literature on modern slavery. The evaluation team, including research and evaluation partner organizations based in Asia, consulted with foundation staff, the donor, and the entire MELU team and its director. A theory-driven case study approach was adopted for the evaluations, and both the baseline study and the evaluations were designed to contribute toward building case studies, each focused on the program's three sectors: garment work, domestic work, and labor migration.

During the inception phase, the team identified multiple risks for the evaluation, along with means for managing those risks, and shared them with the donor, at the donor's request. The donor also had a summative quality assurance process (implemented by a contracted third party) for assessing reports submitted. The evaluation team, however, wanted to lead and participate in the quality assurance process, both by ensuring ongoing and sound formative metaevaluation practice and by creating a structure and means by which to engage the donor with quality assurance as necessary throughout the evaluation. We developed a formative metaevaluative framework for the purpose of quality assurance, fashioned along the lines of Scriven's checklist-based approach.

Method for Tool Development

The development of this quality assurance tool was an iterative process. While we were developing a means of ensuring quality for this particular evaluation, we also gave thought to a design that could be applied to any program evaluation for the purpose of promoting quality in praxis. We decided on a framework in a table format with 10 vertical columns and 10 horizontal rows comprising three phases: design, implementation, and reflection. We intend the three phases to be used during the pre-formative stage to map out in advance the evaluation's critical moments and relevant standards. As in Scriven's (2007b) approach, we developed the checklist to serve as a mnemonic device to anticipate critical moments, identify and detail their associated standards and principles, and determine possible actions to adopt and questions to ask. As an iterative tool, we intended to revisit the checklist during the evaluation, with further reflection devoted to preparation (design), addressing the critical moment (implementation), and consideration of the outcome (reflection). We also intended for the checklist to provide a means for generating learning that may be applied to future evaluations, and a means for documentating decisions made on particular issues. Table 1 details the 10 column headings, including, for each, a definition and an indication of its place within one of the three phases.

Table 1. Column Headings for the Formative Metaevaluation Tool

Column	Heading	Definition	Phase
A	Critical moment	The central issue or moment in the evaluation expected to involve decision-making	Design
B	Program evaluation standards	Identification of standards that may come into play in addressing the critical moment	Design
C	Ethics and principles guidance	Identification of ethics and principles that may serve to guide in addressing the critical moment	Design
D	Degree of extrinsic gravity	The evaluation team's understanding of other stakeholders' areas of concern	Design
E	Degree of intrinsic gravity	The evaluation team's areas of concern and priority	Design
F	Questions	Detailing of questions to ask when and of whom	Design/ Implementation
G	Action	Detailing of actions to take	Design/ Implementation
H	Desired outcome	Detailing of desired outcome to actions taken	Design/ Implementation
I	Outcome realized	Detailing of outcome of F and G	Reflection
J	Observations/ comments	Detailing main points of reflection based on Columns A through I	Reflection

Questions driving the development of the quality assurance tool can be considered at each of the three stages. When planning for implementation, the following questions would all assist in reflection and generate specific moments and actions to attend to during the course of the evaluation: What are the critical moments? What could go wrong? What do I need to be wary of? Listing critical moments and identifying relevant standards and principles can be an iterative process; additional critical moments may be identified through the review of program documentation.

We populated the first column of the tool with the critical moments we identified, then we identified each critical moment's corresponding standards and principles. For this purpose, we used the Program Evaluation Standards (Yarbrough et al, 2010), and the U.K. Department for International Development's ethical guidance for research, evaluation and monitoring activities (DFID, 2019). In the process of reviewing the standards and principles, we identified additional critical moments. We continued this iterative process until a well-rounded group of critical moments, along with their corresponding standards and ethical principles, were listed in Columns A, B and C.

To determine the degree to which a critical moment identified was of gravity, either extrinsically or intrinsically, we employed a simple ranking process, designating each as low, medium, or high. This mindfulness about values and priorities is helpful in planning questions (Column F) and actions (Column G) and ensuring that both stakeholder and evaluation team priorities and concerns are addressed.

In preparing to address the critical moments in the implementation phase, we asked additional questions: What should be done when, how, and with whom? What is the desired outcome? We then developed a metaevaluation timeline for the implementation phase alongside the evaluation timeline, with consideration as to what should or can be done, when, and with whom. The identification of the key moments and where in the evaluation process they should be addressed provided us with a work plan for the formative metaevaluation. We found it useful to plan for implementation and determine what tools may be used, at what times an external perspective might be useful, and how it may be documented.

Table 2. Pertinent Questions by Phase

Design	Implementation	Reflection
• What are the critical moments? • What evaluation standards and principles are relevant?	• What should be done when, how, and with whom? • What is my desired outcome?	• Were desired outcomes achieved? • Were standards met?

In the modern slavery evaluation, we determined the most pertinent critical moments to include: (a) concerns about reconciling the evaluation approach with available resources and (b) the need for credible evidence. The evaluation team designated these as having a high level of intrinsic gravity. Since we were using a case study approach and needed to collect rich and detailed data at multiple project sites within communities in Asia, another concern was how best to work as a transnational and multicultural team in cooperation with the foundation and their grantees to arrive at a rich, descriptive narrative of the context and work undertaken within these communities. We identified relevant sections of the Program Evaluation Standards (Yarbrough et al, 2010): reliable and valid information (A3, A4), sound design and analysis (A6), and justifiable conclusions (A1). Relevant DFID (2019) ethics and principles included (a) the "design and conduct of work" being "sensitive to cultural, socioeconomic, environmental, and political contexts", (b) "harms to individuals and communities" being "minimized and benefits maximized", (c) "risks being identified and mitigating actions being taken" (p. 4).

Based on these standards and principles, we adopted an argumentative and populist evaluation philosophy (Schwandt, 2015) involving a strong multicultural stance to enable joint interpretation of data, mutual learning among all team members, and arrival at a common understanding as a critical piece for collecting and engaging in data analysis. We developed a process to enable reflection as an international and multicultural evaluation team as well as each team member's nurturing of their own awareness and knowledge of "self in dynamic context" (Schwandt, 2003). As a multicultural team evaluating programming implemented in particular contexts in South Asia by organizations originating in other locations, we were particularly focused on program relevance to the local contexts, fit of the program with societal arrangements, and compatibility of the program with goals found within the local contexts (Schwandt, 2015).

For our critical moment of high intrinsic gravity on the need for critical evidence, initial questions we articulated for our tool (Column G) included What are our individual orientations with respect to the program we are examining? And how does our analysis reflect who we are? We identified enabling an aware understanding collectively as a team and challenging each other's perspectives as keys to developing credible case studies reflective of the communities' experiences.

In approaching team data collection and analysis we made a point of engaging with, among other actors, the donors, keeping them abreast of our activities and the norms and values we were abiding by to ensure quality praxis. Our intent in engaging with them was to further build credibility in our findings by conveying the standards we applied.

Working through the tool, we found ourselves valuing Symonette's (2004) framework of (a) mapping the social topography, (b) multilevel dynamic scanning, and (c) cultivating empathic perspective taking, and incorporated it into our work. This framework enabled us, as a diverse team of evaluators, to reflect on and question our interpretation of the data as it related to our understanding of the sociopolitical and sociocultural environment in which we were working. We sought to engage in the social and political structures of how power and privilege may be embedded, and to reflect on our own filters as individuals on a diverse team as we explored the realities of race, class, and gender in the complex problem of modern slavery in South Asia.

Through this process we aimed to achieve a high degree of multicultural validity (Kirkhart, 1995; Symonette, 2004) in our findings and thus meet the identified standards and principles of reliability, validity, cultural sensitivity, and maximized benefits. In doing so we worked to address the extent to which the programs contributed toward sound and equitable development agendas within the communities, and how they might be better oriented to do so, effectively striving to conduct what Ofir (2014) calls "evaluation for development, not evaluation of development" (p. 584).

An area we determined to be of high extrinsic gravity for the evaluation was the evaluation team's ability to maintain independence from baseline to summative evaluation. The donor expressed concern about the evaluation team's inclusion in the MELU and its close association with others working with the foundation and its partners on an ongoing basis. Standards we associated with this concern included valid information (A2), clear and fair assessment (P4), and impartial reporting (A8).

For this critical moment of high extrinsic gravity, initial questions we articulated for our tool (Column G) were How are we experiencing "independence" and "dependence" through our process? And what are some obstacles we see to our independence? Initial actions we identified (Column H) were to track factors that supported or challenged our evaluation team's independence throughout the three-year process, providing a means for reflection on this aspect of the evaluation, and to then raise any concerns or questions with the donor. We were also mindful of the various facets and meanings of independence vis-à-vis all stakeholders, including our donor, and how the donor's view of and concern for impartiality may differ from our own. Before and after each evaluation exercise (baseline, formative, and summative), we prioritized regular communication both between the MELU director and the evaluation team and between the evaluation team and the donor to discuss any particular challenges faced in maintaining independence from all stakeholders, and how we were addressing them.

The final phase of reflection enables a review of the plan's implementation and involves key questions such as What happened? What went well, and what went wrong? And how well were standards and principles met? Reflection may be built in before and after key moments as well as at the end. Additional questions may be developed to guide this reflection. At the writing of this article, the project is on pause because of the coronavirus pandemic, but we expect this reflection to aid in documenting the evaluation process (Yarbrough et al, 2010, E1: Evaluation Documentation) and reflecting more fully on the value of this tool once the project resumes.

Developing the tool for quality praxis enabled our team to be mindful of the critical moments that needed to be carefully tended to in both planning and implementation. The tool provided a means of planning and implementation as well as reflection and learning, and it built a metaevaluative lens into our processes, cueing us to ask important questions about why and how the evaluation would be implemented. This helped us undertake the evaluation with a high degree of validity and usefulness.

Reflection on Process

The theory behind our formative evaluation checklist is simple. It relates to the assumption that evaluation standards and principles are applicable to any given evaluation context, and the fact that there are many standards and principles to choose from for any particular situation. Another, and perhaps more pertinent, assumption is that predicting critical moments and identifying their associated standards and principles and possible actions leads to higher-quality practice in evaluation.

The generic but action-oriented quality of the standards was useful both in linking critical moments to relevant standards and in thinking through possible actions relevant to different circumstances. Review of the standards allowed us to populate the cells in Column B (program evaluation standards), Column F (questions), and Column G (action). In addressing critical moments, we identified the need to gather information and determine expectations held among stakeholders. For example, we identified the need to survey stakeholders about their expectations for the evaluation, and to engage stakeholders in developing a clear rationale for value judgments.

Further, the standards checklist enabled an iterative process of identifying not only relevant standards and possible actions to take, but also additional critical moments. Among these later-identified moments were (a) dissemination of findings, (b) ensuring confidentiality in management of data, and (c) aligning expectations and work plan for the evaluation following approval of the evaluation design report. While these additional moments were not designated as high in intrinsic or extrinsic gravity as compared to those identified earlier, they were relevant to the evaluation, and identifying them contributed to sound, quality evaluation practice.

While the standards were useful for identifying criteria and actions to take, the standards checklist does not make explicit a standard for a given quality of practice or outcome for each critical moment. The use of the Program Evaluation Standards as criteria, not standards, is evident in, for example, the noting of reliable, systematic, and valid information as pertinent to the concern of acquiring enough quality data to contribute toward building the case study for the evaluation. The degree to which data are reliable, systematic, and valid is left to the evaluator to determine, and that determination may be informed by practical experience and knowledge, the phronesis that Schwandt refers to. Wingate's (2009) study on raters' varied uses of the standards in metaevaluation reflects this reality, leading to her conclusion that metaevaluative judgments using the standards are "largely idiosyncratic" (p. 107). While clear determinations of what constitutes goodness are not found in the Program Evaluation Standards, we note and agree with Wingate's (2009) conclusion that the standards serve as a beneficial open-ended guide to reflect on one's practice, in concert with one's own experience and knowledge.

The tool's differentiation between extrinsic and intrinsic degrees of gravity enables an assessment of each incident's level of importance to various stakeholders, aided by the Program Evaluation Standards and the DFID principles. This assessment offers an explicit opportunity to uncover potential value differences between the evaluation team and the client. By addressing areas of concern both extrinsic and intrinsic to the evaluation team, the team gains a clear overview of priorities and concerns among stakeholders and can map out when and how to address them.

The second assumption related to our tool—that documenting critical moments, their associated standards and principles, and possible actions will lead to quality evaluation practice—raises the question of how quality practice is achieved. It may well be the case that all evaluators make unexamined or subconscious decisions as they evaluate. Undertaking such a formative metaevaluative approach may encourage more explicit and detailed thinking, enabling evaluators to, as Schwandt (2003) writes, adopt practical reason and achieve self-awareness in their approaches. Explicit documentation of critical moments can help evaluators explore the nuances of an evaluation and take active steps to ensure quality in practice.

In our own experience, the process enabled preparedness and planning for many contingencies. Ultimately, thinking through the critical moments and striving to maintain standards prepares evaluators to produce more credible and valid findings and to be prepared for summative metaevaluation. Documenting ideas and plans leads to clearer and more effective planning, including for how and when to act, as well as how to respond in contingencies. Further, the tool facilitates stakeholder engagement, enabling preparation for meetings and ensuring stakeholder concerns for quality are met.

The reflective nature of the tool also encourages more nuanced thinking throughout the evaluation process. Documenting outcomes of questions asked and actions taken to assure quality provides an opportunity to reflect on what works and what may not, and to apply those lessons to the same evaluation or to future evaluation work. The documentation may then be used for future reference and reflection.

Ultimately, the tool is intended to further professionalize the evaluation field, enabling evaluators to negotiate the extrinsic and intrinsic merits of our work with commissioners and other stakeholders, and to take the lead and engage with our donors and other stakeholders about how we perform quality praxis.

Conclusion

Metaevaluation may be regarded as intrinsic to what we as evaluators do; quality is often assumed to be baked into our work and processes. Perhaps our competitive advantage, as described by Picciotto (2011), makes evaluators inclined to self-evaluate naturally. Yet most of the evaluation literature is descriptive when it comes to metaevaluation, and how we as evaluators actually metaevaluate or should metaevaluate is not well researched and understood.

Christie's (2003) research tells us that there tends to be a disconnect in the evaluation field between aspirational theories and what is actually done in practice. Her reflection on research on evaluation as a meeting place for theorists and practitioners may facilitate a move from prescriptive theories that are theoretical and deductive to descriptive theories that are experience-based. Formative metaevaluation practice may provide a means for moving in this direction. It may guide deliberate reflection, elevating planning and ensuring quality in one's work as an avenue for bringing in components of our own aspirational theories and ensuring that we as evaluators are practicing to a high professional standard. In effect, formative metaevaluation practice is a means for accountability to one's theory of practice.

Formative metaevaluation enables evaluators to take greater control in deliberately ensuring quality and negotiating quality with commissioners. By taking such control we define quality for our field and make steps to address the imbalance of power in our fee-dependent profession (Picciotto, 2011). As evaluators we also become more mindful and skilled in understanding and applying our industry standards and principles. We promote what Scriven (1994) calls the consumer-oriented view of evaluation. The welfare of the "consumer," the population the program intends to serve, is a program's primary justification, above the concerns of management or other stakeholders, and consumers' welfare is to be prioritized in the evaluation as well.

Evaluators aim to go beyond serving as mere technicians who have the skills to carry out the interests of commissioners. We aim to be mindful of the ethics of evaluation practice, or to have what Schwandt (2015) terms an "ethical disposition" as evaluation professionals. In doing so, we apply phronesis and exercise professional judgement. Formative metaevaluation enables evaluators to more purposefully practice phronesis and thereby contribute toward advancing and strengthening the evaluation profession.

We offer our tool as a first step in our research on formative metaevaluation practice, and we would value the evaluation community's input as we aim to empirically test it. We welcome partnership with others in this effort.

Acknowledgment

The authors would like to acknowledge the helpful comments of the peer reviewers in finalizing this article.

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

This study was supported by the Interdisciplinary PhD in Evaluation (IDPE) program at Western Michigan University (WMU) and The Graduate College at Western Michigan University (WMU).

References

Abbott, A. (1988). *The system of professions*. University of Chicago Press.

American Evaluation Association (2018). Guiding principles. https://www.eval.org/Portals/0/Docs/AEA_289398-18_GuidingPrinciples_Brochure_2.pdf

Christie, C. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. *New Directions for Evaluation*, *97*, 7–35.

Cooksy, L. J., & Caracelli, V. J. (2009). Metaevaluation in practice: Selection and application of criteria. *Journal of MultiDisciplinary Evaluation*, *6*(11), 1–15.

Dahler-Larsen, P. (2019). *Quality: From Plato to performance*. Springer International Publishing Imprint: Palgrave Macmillan.

DFID (2019). *DFID ethical guidance for research, evaluation and monitoring activities*. https://www.gov.uk/government/publications/dfid-ethical-guidance-for-research-evaluation-and-monitoring-activities

Fitzpatrick, J., Sanders, J., & Worthen, B. (2011). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.). Pearson.

Friedson, E. (1983). *Professionalism: The third logic*. University of Chicago Press.

Harnar, M. A., Hillman, J. A., Endres, C. L., & Snow, J. Z. (2020). Internal formative meta-evaluation: Assuring quality in evaluation practice. *American Journal of Evaluation*, *41*(4), 603–613.

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. *New Directions for Evaluation*, *97*, 69–80.

Jacobs, S., & Affrodegon, W. S. (2015). Conducting quality evaluations: Four generations of meta-evaluation. *SpazioFilosofico, 13*, 165–175.

Kirkhart, K. (1995). Seeking multicultural validity: A postcard from the road. *Evaluation Practice*, *16*(1), 1–12.

Lincoln, Y. S. (1985). The ERS standards for program evaluation. *Evaluation and Program Planning*, *8*(3), 251–253.

Ofir, Z. (2013). Strengthening evaluation for development. *American Journal of Evaluation*, *34*(4), 582–586.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.) Sage.

Picciotto, R. (2011). The logic of evaluation professionalism. *Evaluation*, *17*(2), 165–180.

Picciotto, R. (2020). From disenchantment to renewal. *Evaluation*, *26*(1), 49–60.

Sanders, J. R. (1995). Standards and principles. *New Directions for Program Evaluation*, *66*, 47–52.

Schwandt, T. (2003). 'Back to the rough ground!' Beyond theory to practice in evaluation. *Evaluation*, *9*(3), 353–364.

Schwandt, T. (2015). *Evaluation Foundations Revisited: Cultivating a Life of the Mind for Practice*. Stanford University Press.

Scriven, M. (1969). An introduction to meta-evaluation. *Educational Products Report*, *2*(5), 36–38.

Scriven, M. (1991). *Evaluation Thesaurus* (4th ed.). Sage.

Scriven, M. (1994). Evaluation as a discipline. *Studies in Educational Evaluation*, *20*, 147–166.

Scriven, M. (2007a). Key evaluation checklist. https://wmich.edu/evaluation/checklists

Scriven, M. (2007b, unpublished). The logic and methodology of checklists.

Scriven, M. (2009). Meta-evaluation revisited. *Journal of MultiDisciplinary Evaluation*, *6*(11), iii–viii.

Scriven, M. (2012). Formative, preformative, and proformative evaluation. *Journal of MultiDisciplinary Evaluation*, *8*(18), 58–61.

Société Canadienne d'Evaluation (SCE) (2011). *The program evaluation standards: A guide for evaluators and evaluation users*. Sage.

Societe Suisse d'Evaluation (SEVAL) (2016). Evaluation standards of the Swiss Evaluation Society. https://www.seval.ch/app/uploads/2018/08/SEVAL-Standards-2016_e.pdf

Stufflebeam, D. L. (1999). Program evaluation metaevaluation checklist. https://wmich.edu/evaluation/checklists

Stufflebeam, D. L. (2001a). The metaevaluation imperative. *American Journal of Evaluation*, *22*, 183–209.

Stufflebeam, D. L. (2001b). Evaluation checklists: Practical tools for guiding and judging evaluations. *American Journal of Evaluation*, *22*(1), 71–79.

Stufflebeam, D. L., & Shinkfield, A. (2007). *Evaluation theory, models, and applications*. John Wiley.

Stufflebeam, D. L. (2011) Meta-evaluation. *Journal of MultiDisciplinary Evaluation*, *7*(15), 99–158.

Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models, and applications* (2nd ed.). Josse-Bass.

Symonette, H. (2004). Walking pathways toward becoming a culturally competent evaluator: Boundaries, borderlands and border crossings. *New Directions for Evaluation*, 102, 95–109.

United Nations Evaluation Group (2020). Ethical guidelines for evaluation. http://www.unevaluation.org/document/detail/2866.

Wingate, L. A. (2009). *The program evaluation standards applied for metaevaluation purposes: Investigating interrater reliability and implications for use.* [Unpublished doctoral dissertation]. Western Michigan University.

Yanagi, M., & Leach, B. (1972). *The unknown craftsman: A Japanese insight into beauty* (1st ed.; B. Leach, Adapt.) Kodansha International.

Yarbrough, D. B., Shula, L. M., Hopson, R. K., & Caruthers, F. A. (2010). *The Program Evaluation Standards: A guide for evaluators and evaluation users* (3rd. ed). Thousand Oaks, CA: Corwin Press.