Journal of MultiDisciplinary Evaluation
Volume 20, Issue 47, 2024

JMDE
Journal of MultiDisciplinary Evaluation

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

Revisiting and Expanding Scriven's Fallacies About Formative and Summative Evaluation

Janet Clinton and John Hattie[1]
University of Melbourne

The Origins of Formative and Summative Evaluation

While evaluation has been the hallmark of human endeavor for millennia, the discipline of evaluation is relatively new. Michael Scriven was one of the discipline's founding members, inventing much of its infrastructure, promoting its rigor and methods, and infusing it with his philosophical acumen. He shaped many of the roots of the "evaluation theory tree." Along with Tyler, Stake, and Stufflebeam, Scriven took part in many core debates, and none was so rich as the debate about the purpose of evaluation. Scriven argued that evaluation is a discipline, the alpha discipline, with a well-defined subject matter, a logical structure, and many application areas. Most importantly, he grounded evaluation in the logic of valuing (Scriven, 2012).

Within this thinking, the concepts of formative and summative evaluation emerged. These concepts derived from a battle between Scriven and Lee Cronbach, with Cronbach preferring an emphasis on formative evaluation and Michael arguing that summative evaluation was of similar, high value (depending on context). In later years, like many evaluators Scriven argued that the impact of evaluation relates to its formative influence on the evaluand regardless of the process, but he continued to promote claims about the importance of summative evaluation.

Within the formative and summative debate, Cronbach and Suppes (1969) distinguished between conclusion-oriented evaluation and decision-oriented investigations. Conclusion-oriented studies take direction from the investigator's commitments and hunches to conceptualize and understand the chosen phenomenon and to freely reframe the questions as the study progresses, and are preferred by academics to advance their careers. Decision-oriented studies aim to provide information a decision maker wants, are typically commissioned studies, and have major constraints regarding the freedom to modify the questions or "wander down interesting bypaths or to burrow into deeper questions" (p. 21). In Cronbach's (1982) view, the priority of conclusion and decision evaluations was to serve the decision maker, where there is consensus about the goals and program operation and the evaluator is informed what to look for; occasionally the commissioning agent may "reduce the evaluator to a technician by setting forth the questions to be answered and asking him simply to apply his skills of sampling measurement, and statistical analysis" (p. 6). Critically, "It is not the evaluator's task to determine on his own whether a program is worthwhile or what action should be taken" (p. 7–8). The argument is that the "evaluator begins to educate his clientele as soon as he begins to interact with its members … they feed into the planning process directly and also through the reactions they elicit from those with whom the evaluation talks" (p. 10). Cronbach (1964) argued that "evaluation, used to improve the course while it is still fluid, contributes more to the improvement of education than evaluation

[1] Thanks to Dylan William for helpful improvements.

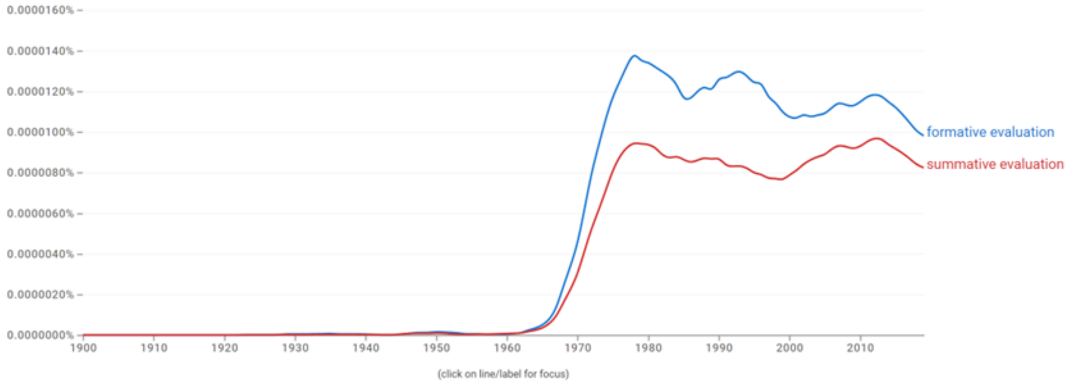
used to appraise a product already placed on the market" (p. 236).

Scriven (1967) began his response by distinguishing between the roles and goals of evaluation. The *goals* are many but primarily relate to merit, worth, and value (How well does ..., Does it perform better than ...), whereas the *roles* of evaluation are remarkedly varied. We cannot, he argued, reduce evaluation to the roles it can take, but to the goal of establishing merit, worth, and significance. For Scriven (2013), evaluation means a "cognitive process or act of evaluation, that is, the determining or asserting of a claim about the merit (a.k.a. roughly speaking, quality), worth (a.k.a in one sense, value), or significance (a.k.a. approximately, importance) of some entity" (p. 13).

Scriven (1991) disagreed with the assertion that formative evaluation is of greater importance than summative evaluation. He noted that calling in an evaluator to perform a final summative evaluation is most worthwhile, and that it "seems a little excessive to refer to this as simply a 'menial role,' as Cronbach does" (p. 42). He countered Cronbach's (1963) claim that "evaluation, used to improve the course while it is still fluid, contributes more to the improvement of evaluation than evaluation used to appraise a product already placed on the market" (p. 236). Formative evaluation judges the worthwhileness of a program, process, or product *during* its development, and summative evaluation makes such judgments nearer to or at the end of the development process. For example, in educational evaluation, formative evaluation is "simply outcome evaluation of an intermediate stage in the development of the teaching instrument" (Scriven, 1967, p. 51), and its role is to discover deficiencies and successes in the intermediate versions of a new curriculum. It asks not how well the course achieves its goals but how good the course is.

Since these debates, the terms "formative" and "summative" have been used extensively, pirated, misused, and spread across many disciplines. Indeed, the terms have been popularized into everyday language. It is possible to trace the terms "formative evaluation" and "summative evaluation" from Scriven's introduction until today (Figure 1).

Figure 1. NGram of the Frequency of Use of the Terms "Formative" and "Summative" from 1900 to 2020



We contend that the terms have lost their original meaning and power, particularly in the education discipline, where the concepts and methods "formative" and "summative" are geared far too much toward assessment. Our major argument is that Michael's original descriptions and interpretations of these evaluation processes remain critical bedrocks for the optimal use of these terms, and it is worth reverting to using the terms "formative" and "summative" as he intended rather than conflating the ideas with assessment.

Identifying Fallacies

Twenty-five years later, Scriven (1990) updated his formative and summative evaluation thinking. He started by noting Bob Stake's analogy: When the cook tastes the soup, it is formative; when the guests taste the soup, it is summative evaluation. He noted continual confusion about the origin in the uses of these concepts and many major errors that have crept into using the terms. He identified several fallacies, and we have noted further fallacies, leading to some major recommendations for moving forward.

The first fallacy is that *formative and summative evaluation are intrinsically different types of evaluation*. No, they are not different types, but they serve a specific purpose—formative evaluation is designed, done, and delivered with the aim of making improvements to the evaluand. Summative evaluation is done for, or by, any decision makers who need evaluative conclusions for any reason *other than* conceptual development. So improvement or status; during or after; in-flight or landing. It is less formative *versus* summative but often formative *and* summative: Scriven (1993) argued that "good formative may begin by being mild-mannered, but not too far into the development process it needs to remember that one of its functions is to provide 'early-warning summative' to the project staff" (p. 210).

Scriven distinguished five levels of formative: (0) improvements by the author/inventor to a program or entity; (1) in-house critiques by colleagues or employees not in the development group; (2) field trials supported by representatives of the development team; (3) hands-off field trials at remote sites by supposedly typical users working on their own in their usual environments (the classic beta-testing in software development); and (4) full-scale commissioned evaluations by external expert reviewers who run systematic experiments with end users. The key concept is that in "light of these processes, or some of them, the product is (or is not) finally revised and released, and summative evaluation begins" (1993, p. 3). So, the distinction is that the purpose is formative during and summative at a key milestone throughout the entity's life course (Schwandt, 2018); they can complement each other and sometimes work in parallel. It is not either/or, but when.

Any hint that formative is less rigorous than summative must be dismissed. Thus, Scriven's second fallacy is that *formative evaluation can be a much more informal process than summative evaluation*. To the contrary, if we are to make optimal improvement decisions based on formative evaluations, then surely the standards for formative evaluation need to be very high, as mistakes are expensive. "Doing formative evaluation any less rigorously than a good summative evaluation simply undermines the accuracy of the midcourse corrections, which is all too likely to send the mission in the wrong direction" (Scriven, 1991 p. 7). Furthermore, nothing in the formative stage suggests that it needs to be completed only internally and does not warrant the involvement of external evaluators.

The third fallacy, that *formative is more worthwhile than summative*, is derived from the influence of Cronbach (1963), who argued for a shift away from summative to formative evaluation, and also away from comparative evaluation to evaluation in isolation. As noted above, Cronbach was not a fan of summative evaluation, preferring a "kinder, gentler" method, arguing that evaluations need to be used almost entirely in a formative manner and that "these handy terms [formative and summative] are not adequate for today's discussion" (Cronbach, 1964, p. 236)—which, of course, has been contradicted by the extensive use of the terms since they were invented. A major reason for raising the relative merits of formative and summative evaluation was Scriven's argument that "summative evaluation is very powerful, rightly or wrongly" (1991, p. 5). For example, "Who wants their children taught to read using a method which is only half as effective and no more fun than another program at equal cost?" (1991, p. 52). He argued that the summative evaluation of tests, essays, texts, and teachers is critical.

Scriven's fourth fallacy is that *formative evaluation does not lead to an overall rating* but primarily includes recommendations for improvement or causal explanations of performance. Scriven uses the example of essay grading that may lead to an overall judgment to promote formative improvements by the student—a judgment that, in many instances, this can be more reliable and dependable, with more feedback power than specific or analytic sets of scores related to some rubric. Hence, the worth of evaluation is its power to inform and influence change.

The fifth fallacy is the belief that *the evaluator's duty is to give the evaluand the facts and then let them interpret them according to their values*. This leads to the evaluator doing everything except what the purpose of 'evaluating something' is: making decisions of merit, worth or significance. It is the evaluator's job to call it. This is where there is often a slip, when the person completing the evaluation also interprets the evaluation, whereas critique, interpretations by others (internal or external), and triangulation become critical. Ensuring that an evaluative judgment is made on the quality of the work and the evaluative process is vital. Much of our work from Visible Learning (Hattie, 2008, 2023) is premised on this notion that a teacher's selfreflection can be as much of a negative as a positive influence, as we (humans) have remarkable biases to see the world through our eyes, and take credit for our successes and blame others (students, leaders, resources, context) for failures. What makes the greatest difference is when there is the triangulation of evidence (from test scores, from artifacts of student work, from progress, from students, as well as from teacher noticing and judgments), and when this is part of a collective interpretation—seeking multiple viewpoints, seeking evidence you may be wrong in your judgment, seeking alternative next best steps. Our evaluands have the right to seek a "second opinion"—for both formative and summative interpretations.

The sixth fallacy maintains that *in formative evaluation it is only necessary to point out various respects in which improvement is needed*. Indeed, Scriven argues, this is the "most dangerous of our agenda of mistakes" (1991, p. 12). Scriven is trying to counter Cronbach's "friendly formative" image where there are no threatening overall conclusions, no deep diagnosis, and no recommendations for improvement. The worst treachery, Scriven claims, is when there is resistance to producing or circulating negative summative evaluations, or deliberate ignorance of improvement suggestions that point to negative attributes (insufficient dosage, quality of implementor and implementation, etc.). Over the course of an evaluation, structure and logic must be explored and the truth brought to the fore.

The seventh fallacy is that *formative evaluation must cause or lead to improvement*. As one example, Dolin et al. (2018) argued that the "'constructive' use of formative assessment hinges on the ability of the teacher (or another provider of feedback) actually to give recommendations that are relevant and effective for improvement" (p. 59). But the mechanics and responsibilities for improvement are beyond the evaluator's role. A formative evaluation should be completed in a way that facilitates enhanced development by the provision of evaluative judgments, but evaluators are not responsible for whether any recommendations are implemented or not. Too often, accountability systems not only demand formative evaluations but also ask the accountability gurus to lead the improvements. This leads to primarily looking at what the evaluand can improve (e.g., principals evaluate teachers in terms of their skills and resources for improvement and thus can bias the evaluation by not evaluating many aspects outside the evaluator's control to implement). From an evaluation logic perspective, it is primarily about the explanation.

Scriven firmly believed there were many forms of evaluative thinking, based on creativity, reason, logic, ethics, and seeking multiple and contrary views. Such interpretative activities within an evaluation lead to decisions about merit, worth, and significance to (in formative evaluation) improve what's being evaluated and (in summative evaluation) document achievements or outcomes of the program or processes. The results of both activities can be acted out. Some of the problems of formative evaluation occur when there is no opportunity for follow-up; when the quality of the formative evaluation information is not valid, checked, or triangulated; when there is little checking on whether recipients hear, understand, and can action the formative information; and when the recipients are merely asked to repeat the task with the same form of engagement that led to the issues.

The eighth fallacy is that *quality evaluation of a particular content area requires the evaluator to have demonstrated skill in performing that job or in that area*. This is, argues Scriven, putting the formative cart before the summative horse. Sometimes, for example, the best evaluators of teachers are not necessarily those immersed in teaching or in leading the teachers, but external evaluators who come to the task from a less internally biased viewpoint.

Scriven concluded his 1991 paper outlining these fallacies by stating that his aim was a "defense of summative evaluation and global evaluation against various attacks" (p. 62). He saw formative evaluation as becoming too dominant and wanted to balance the equation by arguing the merits of the various formative and summative evaluation roles. His bottom line is that "doing good evaluation" is the most critical claim, whether the purpose is formative or summative.

Identifying More Recent Fallacies

The ninth fallacy is that *there are such concepts as formative and summative testing*. Michael was not at all impressed when Bloom et al. (1971) applied these terms to education and learning with the release of their book Handbook on Formative and Summative Evaluation of Student Learning. There was little acknowledgment in the book of Scriven's definitions of the terms, and Michael was critically disturbed at the gross misinterpretations that described tests as either formative or summative. We do note that Bloom et al. did not use the term "testing" in the book's title, and that they did claim, "We have borrowed the term Formative Evaluation from Scriven (1967) to refer to these diagnostic-progress tests" (p. 9). But the damage was done throughout the book by linking the two terms to tests and testing. Bloom et al. described "formative assessment" as assessment that "aids both the teaching and learning process ... while they are still fluid and susceptible to modification" (p. 20), and they used formative evaluation within their mastery teaching model such that teachers organized what they wanted students to learn and then assigned "formative tests to each of these units" (Bloom 1968, p. 20). "By formative evaluation we mean evaluation by brief tests used by teachers and students as aids in the learning process" (Bloom et al., 1971, p. 48). This, they argued, improves the student performance in any summative (end-of-course) assessment. Hence, the claim was for the power of formative assessments to improve learning, which then could be measured by summative assessments—beginning the divide between the two notions and (falsely) yoking formative and summative evaluation to testing.

Many have followed Bloom's lead. Cizek (1996; Cizek et al., 2019), for example, applauded that Bloom changed the focus from formative evaluation to formative assessment, as this had a "profound impact on the field of assessment" by moving from the focus in evaluation on "ascribing worth or merit to the results of an information-gathering procedure (such as assigning grades on a test" (Cizek et al., 2019, p. 6). Dunn and Mulvenon (2019) defined summative assessments as those designed to determine academic development after a set unit of study (i.e., assessment of learning), and formative assessments as desired to monitor student progress during the learning process (i.e., assessment for learning). Guskey (2010), one of Bloom's students and a major contributor to the assessment literature, noted that Bloom's interest was in explaining variation in student achievement, and hence that it was important that teachers increased the variation in their instructional practices. This could be helped, Guskey argued, if teachers used their classroom assessments as formative learning tools followed by diagnostic, feedback, and corrective procedures.

The cat was out of the bag, as so many now refer to assessment as formative or summative, whereas any test can have formative or summative interpretations. It is the purpose of the interpretation that differs, not the tests. Further, there is a false inference that "formative and summative tests" have different psychometric properties. The core notion is that any test can be interpreted formatively or summatively, and that it is misleading to speak of "formative and summative assessments." Expunging this concept would greatly help improvement in schools, as educators would stop seeing the tests as formative or summative. This can lead to simply administering assessments during or at the end of instruction and sadly not making the necessary evaluative interpretations to improve learning. Formative and summative tests can easily lead to data collection during or at the end of instruction and not engaging in the more critical formative or summative evaluations based on these data. What is critical are the interpretations, the quality of the feedback provided (to teachers and to students), and the opportunities provided to improve.

Relating to this fallacy is the tenth fallacy: that *formative is good and summative is bad*. The major moves of accountability over the past decades have introduced public naming, shaming, and blaming of schools for not making "every student above average." Cronbach wanted to disparage summative interpretations because they too readily led to such negative effects. Further, the rise of the worldwide accountability-testing movement has also led to summative still being seen as bad, sad, and mad. Critics therefore argue that such summative assessments not only are of little use but also can lead to negative impacts.

The claim that formative is good, has also led many test development companies to relabel their test products as "formative assessments" (Popham, 2006). Many debates were had during the development of the New Zealand assessment system (https://e-asttle.tki.org.nz/), and many times Michael (then professor of evaluation at the University of Auckland, the first to hold that title) was asked to help argue that the assessments could be used for formative and summative purposes and resist the pleas of government and the teacher groups to call the system (e-asTTle) we developed "formative assessment" (Hattie et al., 2005). Teachers wanted "formative assessments" but initially used e-asTTle tests more for summative purposes, and much professional learning was undertaken to demonstrate the different purposes, values, and interpretations of both purposes of assessment and how these interpretations could be derived from the same tests.

Lau (2016) noted the slogan in George Orwell's *Animal Farm*, "Four legs good, two legs bad," which, in that story, was chanted to the point that the underlying principles were forgotten, and we note how similarly the mantra within evaluation seems to be developing: formative good, summative bad. Lau argues that Scriven (and Bloom) wanted to see roles for both formative and summative interpretations, but due to Bloom's emphasis on different types of tests for formative and summative and his claims prioritizing "formative assessment as a way to improve students' summative performance" (Lau, 2016, p. 512), the link was broken and thence lost. She traces the influence and pressure of external accountability assessment on the image of formative as good and summative as bad.

Lau also claims a major reason for promoting formative assessments, was the influence of Black and Wiliam (2003) pushing "formative assessment" as desirable. She notes the many claims that "summative assessment" can have negative impacts on student motivation, the increasing literature claiming "summative tests" are the "old" model and should be abandoned. Her claim was that these two claims have mistakenly become "direct antitheses of each other" (e.g., Gipps, 1994; Hager & Butler, 1996; Shepard, 2000). There are many references to "new" assessment models, moving from labeling to promoting understanding and growth, more Vygotsky and less Terman and Skinner, pleas for multi-, not uni-dimensionality, arguments supporting thinking and reasoning, not psychometrics, and many other dualities. Formative became Cinderella, and summative became the wicked stepmother.

The eleventh fallacy is that *formative is primarily about comments and information, and summative is primarily about grades or scores*. The claim in education is that formative assessments should not be graded but only include comments, and summative should be used to assign grades. The typical argument for this fallacy is that grades negatively affect student learning (McMorran et al., 2017), as if formative comments are always informative and not sometimes vacuous, personal, or demeaning. This notion was pushed by Bloom and colleagues' argument that "formative assessments" should only be deemed "Mastery" or "Not Mastery" (Bloom, 1968; Bloom et al., 1971).

Following these good/bad claims, the twelfth fallacy is that *interventions based on formative assessment are uniquely good and interventions based on summative are not*. The most prominent "formative assessment intervention" is by Black and Wiliam (1998a). Their seminal article argued that formative assessment "does not have a tightly defined and widely accepted meaning" and "the term 'formative assessment' is not common in the assessment literature" (p. 53). On the contrary, Scriven was clear about the definition and meaning of the terms, but by adding "assessment" they became muddled by others. Black and Wiliam refer to formative assessment as those activities undertaken by teachers and their students which provide information to be used as feedback to modify the teaching and learning activities. Note the broadening from assessment to "those activities." In a related article they argued that "formative assessment is an essential component of classroom work" and that they "they know of no other way of raising standards for which such a strong prima facie case can be made" (Black & Wiliam, 1998b, p. 8).

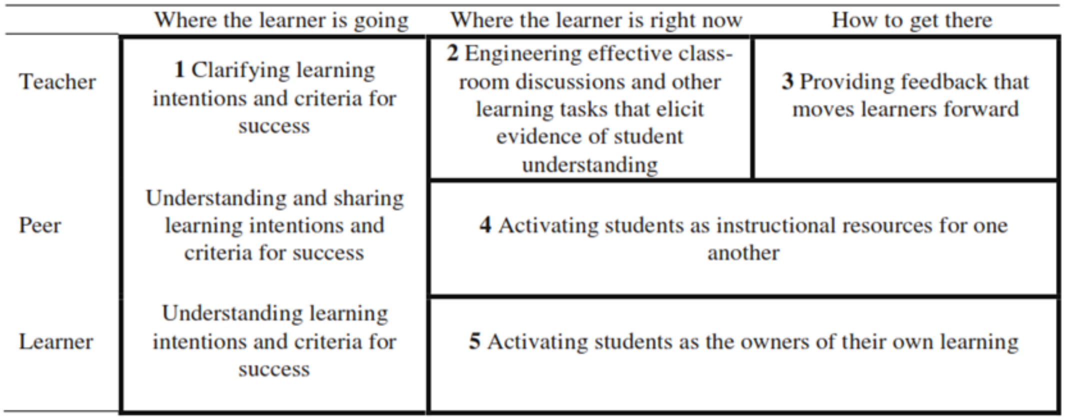
Despite their misleading use of "formative and summative assessment," throughout these two articles and many subsequent resources, we argue that Black and Wiliam have used "formative and summative assessment" to refer more to evaluation than to assessment. For example, they claimed (2018) that "the same assessment instrument, and even the same assessment outcomes, could be used both formatively and summatively" (p. 553). Still, many have misinterpreted their work as if the terms related to "assessment", which is not helped by their consistent use of "formative assessment." This misinterpretation of the terms continued to be replicated as, for example, Black and Wiliam (2009) proposed that assessment is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited (p. 6). No, the assessments are not formative; it is the evaluation interpretations that can serve as formative.

Wiliam later argued that "the biggest mistake that Paul and I made was calling this stuff 'assessment' … because when you use the word assessment, people think about tests and exams" (Booth, 2017, p. 2, and he suggested that their model probably should have been called something like "responsive teaching" (Wiliam, 2013). He also notes the terms became pirated by governments to call their accountability models formative assessment, which not only sullied the

Black and Wiliam big ideas but also caused governments to claim formative assessment did not work.

The value of the model developed by Wiliam is that it is very extensive and has moved far beyond "assessment" to include many aspects of teaching and learning—clarifying, sharing, and understanding learning intentions and criteria for success; effective classroom discussions and questions; providing feedback that moves learners forward; activating students as instructional resources for one another; and activating students as the owners of their own learning (see Figure 2). In our terms, they have moved closer to asking how teachers think, interpret, and evaluate; how they critique and seek evidence of impact; what they consider are their intentions and goals; and how they engage in diagnoses, checking, and monitoring—all aspects of evaluative thinking, which includes both formative and summative interpretations and decisions (Hattie, 2023). Wiliam's emphasis on "next steps in instruction" can come (as he claims) from both formative and summative decision-making. All these steps could be accomplished without any assessment. However, we would argue that including assessments can provide critical information, teacher noticing, student work artifacts, and students' voices about their progress and attainment.

Figure 2. Black and Wiliam's (2018) Depiction of Aspects of Formative Assessment



	Where the learner is going	Where the learner is right now	How to get there
Teacher	**1** Clarifying learning intentions and criteria for success	**2** Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding	**3** Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	**4** Activating students as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	**5** Activating students as the owners of their own learning	

Note. From Classroom assessment and pedagogy, by P. Black & D. Wiliam (2018). *Assessment in Education: Principles, Policy & Practice*, *25*(6), p. 560. doi.org/10.1080/0969594X.2018.1441807

The thirteenth fallacy is that *formative and summative primarily are the responsibility of evaluators*. This ignores the powerful contribution of participants (e.g., students) engaging in formative and summative evaluation. Dolin et al. (2018) claimed, "For some time, it has been recognized that learners have an active role in constructing their understanding; it is not something that can be received ready-made from others" (p. 59). Black and Wiliam (1998b) also emphasized the importance of students needing to be actively involved. Clinton et al., (2023) have, for example, developed the notion of student assessment capabilities, student collective efficacy, and students driving their instruction such that they seek, receive, interpret and engage in formative and summative evaluative thinking (see also Paproth et al., 2023).

The fourteenth fallacy is that *formative and summative evaluation has been confused with process and product evaluation*. Scriven (1996) argued that formative evaluation could be completed by only reporting on outcomes (e.g., in teaching people how to improve their accuracy in firing a pistol by talking in the language of "four o'clock in the nine ring"). Similarly, summative evaluation can be largely or entirely process evaluation (e.g., when sustained abuse of patients and funding is reported as grounds for closure of a nursing home). Scriven (in his reply to Chen's 1996 critique) reiterated his major point: the difference between formative and summative evaluation is "not intrinsic, it's contextual—mainly a matter of the use to which evaluation is put … it is a difference of roles" (Scriven, 1996, p. 153). Returning to the soup, the opinion of the Michelin representative visiting a restaurant may be summative for the Michelin guide and its readers but formative for the chef. Further, evaluation developed for formative purposes can be de facto summative, “for example, when time runs out for a budget decision and the formative evaluation is all that is available to the decision-maker” (Scriven, 1996, p. 153).

The fifteenth fallacy is that *there are but two roles for evaluation: formative and summative*. Scriven has always talked about ascriptive evaluation and occasionally conceded developmental evaluation could be an addition. “I was never very keen on the idea that formative and summative was all there was” (Scriven, 2010, p. 10). Ascriptive evaluation is undertaken for the sake of finding out what the best is. “Nobody’s going to make a decision to disseminate or not, nobody’s looking for ways to improve, but they just want to know the answer” to what is the best (p. 28).

Conclusions

Formative evaluation relates to improvement, and summative relates to status, and both formative and summative interpretations are context dependent: Scriven (1996) used the example of a book reviewer providing summative evaluation of a first edition. If the book goes to a second edition, then this review could be good formative evaluative evidence for making changes. The distinction is improvement or status; during or after; in-flight or landing; and there needs to be less formative *vs.* summative and more formative *and* summative.

In the same way that the cook’s goal is to make the best soup possible for the guests, the cook knows that summative evaluation is coming, making the quality of the formative tasting all the more critical. Poor soup for the guests is pretty powerful evidence of poor formative evaluation in the cooking. If an institution or program has poor summative evaluations in place then it is unlikely they will have the inclination, purpose, or wherewithal to be concerned with formative interpretations throughout the delivery of the program. Too much reliance on tasting the soup, however, may lead to inattention to the goals—such as making soup that is cold when the guests arrive or forgetting to attend to the other elements of the meal.

It is worth noting Scriven’s other claims about formative and summative assessment. First, for example, sometimes we do not necessarily know, when administering an evaluation in a classroom, whether its purpose is formative or summative. For example, an evaluator may consider evaluating the course of teaching by using an end-of-course test but then find that remedial work is necessary, making the test interpretations more formative than intended. Second, formative interpretations can lead to claims that there is already sufficient evidence of program impact, and thus no more dosage is needed. Thus, formative interpretations can become summative interpretations. Third, an emphasis on “improvement,” as if *any* evidence of improvement is necessarily a sign of progress, is a weak yardstick indeed. Claiming there is evidence of “improvement” may be trivial. Too often in education, the goalposts are set so low that the focus turns to satisfaction that there is any progress, and we do not question whether the progress goal was high and appropriately challenging. Hence the need for another of Scriven’s (1973) contributions, goal-free evaluation.

Scriven (1991) argued that when formative evaluation is understood, its primary purpose is to stimulate summative evaluation. Undertaking formative evaluation any less rigorously than summative evaluation undermines the accuracy of any mid-course corrections, which is too likely to send the mission in the wrong direction. Contrary to popular utterance, formative evaluations need to conducted with a high degree of rigor - diagnosis and corrections need to be excellent.

A major problem is that the terms “formative” and “summative” suffer from a concordance with the terms used for assessment. ChatGPT’s answer to the question “What is the difference between formative and summative?” typifies this common usage: It immediately ties formative and summative to assessment (even though the probative question did not ask for this marriage), seeing them as “types” more than roles; it does get the timing correct, but it incorrectly assigns less rigor to formative and more formality to summative. It inappropriately implies summative has little to say about the learning process, but does laudably conclude that both formative and summative (which it repeats are types of assessments) play important roles.

It is time to abandon, expunge, and obliterate the notion of formative and summative assessment and go back to the evaluation foundations. The tie-in with assessment has done major disservice to the original Scriven notions and has led to too much emphasis on the assessments and too little on the timing and quality of interpretative information. It is critical to emphasize the quality of the evaluative thinking in formative, summative, ascriptive, and development roles. We tend to love dichotomies, but there are many roles in evaluation: honest and dishonest roles, explicit and improvement roles, political and professional roles, instrumental payoff and conceptual payoff roles (Scriven, 1996). Relating formative and summative to different implementations of evaluative thinking puts the onus back on adjudging the merit, worth, and significance—which for Scriven is the raison d'être of evaluation.

Michael was more than formative and summative. He was a prolific author [and we have started a database of all Scriven's articles, https://digitised-collections.unimelb.edu.au/collections/d646bc72-9ea5-5b0f-bc53-6b3df7d554c5?spc.sf=dc.date.available&spc.sd=DESC], and he embodied the heart, soul, and mind of evaluation. More than this, he was our friend, mentor, uncle to our three boys, and colleague for over 40 years.

Michael noted:

> The formative/summative distinction caught on like the Black Death, albeit with more amusing consequences. The widespread adoption of the terminology served its author as one of those nice things stored on the top shelf in the intellectual attic that one can pull out on a gloomy day to cheer oneself up, reflecting that one has, after all, made at least a nominal contribution to the discipline. (1996, p. 28)

More than nominal, our friend; a profound contribution of much merit, worth, and significance.

References

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy &* Practice, *5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. Granada Learning.

Black, P., & Wiliam, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal*, *29*(5), 623–637.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*, 5–31.

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi delta kappan*, *92*(1), 81-90.

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment (UCLA-CSIEP)*, *1(2)*, 1–12.

Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw-Hill

Booth, N. (2017, July 9). *What is formative assessment, why hasn't it worked in schools, and how can we make it better in the classroom?* Impact. https://my.chartered.college/impact_article/what-is-formative-assessment-why-hasnt-it-worked-in-schools-and-how-can-we-make-it-better-in-the-classroom/

Chen, H. T. (1996). A comprehensive typology for program evaluation. *Evaluation Practice*, *17*(2), 121–130.

Cizek, G. J. (1996). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of classroom assessment* (pp. 1–32). Academic Press.

Cizek, G. J., Andrade, H. L., & Bennett, R. E. (2019). Formative assessment: History, definition and progress. In H. L. Andrade, R. E. Bennett, & G. C. Cizek (Eds.), *Handbook of formative assessment in the disciplines*. Routledge.

Clinton, J. M., Aston, R., & Paproth, H. (2023, November). *Enhancing teaching practices: The power of evaluative thinking for teachers, school and system leaders* [Symposium]. Australia Association for Research in Education, Melbourne, Australia.

Cronbach, L.J. (1963). Evaluation for course improvement. In R.W. heath (Ed.), *New curricula*. Harper & Row.

Cronbach, L.J. (1964). Learning research and curriculum development. *Journal of Research in Science Teaching*, *2*(3), 204-207.

Cronbach, L.J. (1983). *Desiging evauations of educational and social programs.* Jossey-Bass.

Cronbach, L. J., & Suppes, P. (1969). *Research for tomorrow's schools: Disciplined inquiry for education*. McMillan.

Cronbach, L. J., & Suppes, P. (1969). *Research for tomorrow's schools: Disciplined inquiry for education*. Macmillan.

Dolin, J., Black, P., Harlen, W., & Tiberghien, A. (2018). Exploring relations between formative and summative assessment. In J. Dolin & R. Evans (Eds.), *Transforming assessment: Through an interplay between practice, research and policy* (pp. 53–80). Springer.

Dunn, K. E., & Mulvenon, S. W. (2019). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment, Research, and Evaluation*, *14*(1), 7.

Gipps, C. (1994). Developments in educational assessment: What makes a good test?. *Assessment in Education: Principles, Policy & Practice*, *1*(3), 283–292.

Guskey, T. R. (2010). Formative assessment: The contributions of Benjamin S. Bloom. In H. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 106–124). Routledge.

Hager, P., & Butler, J. (1996). Two models of educational assessment. *Assessment & Evaluation in Higher Education*, *21*(4), 367–378.

Hattie, J. A. C., Brown, G. T., & Keegan, P. (2005). A national teacher-managed, curriculum-based assessment system: Assessment Tools for Teaching & Learning (asTTle). *International Journal of Learning*, *10*, 770–778.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J. (2023). *Visible learning: The sequel: A synthesis of over 2,100 meta-analyses relating to achievement*. Routledge.

Lau, A. M. S. (2016). ‘Formative good, summative bad?’—A review of the dichotomy in assessment literature. *Journal of Further and Higher Education*, *40*(4), 509–525.

McMorran, C., Ragupathi, K., & Luo, S. (2017). Assessment and learning without grades? Motivations and concerns with implementing gradeless learning in higher education. *Assessment & Evaluation in Higher Education*, *42*(3), 361-377.

Paproth, H., Clinton, J. M., & Aston, R. (2023). The role of evaluative thinking in the success of schools as community hubs. In B. Cleveland, S. Backhouse, P. Chandler, I. McShane, J. M. Clinton, & C. Newton (Eds.), *Schools as Community Hubs*. Springer.

Popham, W. J. (2006, September). Defining and enhancing formative assessment [Paper presentation]. Annual Large-Scale Assessment Conference of the Council of Chief State School Officers, San Francisco, CA, United States.

Schwandt, T. A. (2018). Evaluative thinking as a collaborative social practice: The case of boundary judgment making. *New Directions for Evaluation*, *158*, 125–137.

Scriven., M. (1967). The methodology of evaluation. In R. W. Tyler (ed.),Perspective of Curriculum Evaluation, American Educational Research Association (AERA), Monograph of Curriculum Evaluation, No. 1., Chicago: Read Mc. Nally.

Scriven, M. (1973). Goal-free evaluation. In E. R. House (Ed.), *School evaluation: The politics and process* (pp. 319–328). McCutchan.

Scriven, M. (1991). Chapter II: Beyond formative and summative evaluation. *Teachers College Record*, *92*(6), 19–64.

Scriven, M. (1993). The nature of evaluation. *New Directions for Program Evaluation*, *58*, 5.

Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice*, *17(2),* 151–161.

Scriven, M. (2010). In Donaldson, S. I., Patton, M. Q., Fetterman, D., & Scriven, M. (2010). The 2009 Claremont debates: The promise and pitfalls of utilization-focused and empowerment evaluation. *Journal of Multidisciplinary Evaluation*, *6*(13), 15–57.

Scriven, M. (2012). The logic of valuing. *New Directions for Evaluation*, *2012*(133), 17-28.

Scriven, M. (2013). Evaluation as a paradigm for educational research. In E. R. House (Ed.), *New directions in educational evaluation* (pp. 53–67). Routledge.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher*, *29*(7), 4–14.

Wiliam, D. [@dylanwiliam]. (2013, October 23). *Example of a really big mistake: Calling formative assessment formative assessment and not something like “responsive teaching”*. X [formerly Twitter]. https://twitter.com/dylanwiliam/status/393045049337847808.