Why Formative Assessment is Always both Domain-general and Domain-specific and What Matters is the Balance between the Two

Dylan Wiliam

Over the last 50 years, evidence has mounted that assessment can be used to improve the quality of instruction as well as measure its effects. While there is some debate about the magnitude of that improvement, there appears to be increasing consensus that the use of classroom formative assessment needs to be part of any attempt to improve the quality of schools and teachers, and to increase student achievement. As a result of this consensus, attention has shifted to consideration of how teachers can best be supported in developing their use of classroom formative assessment. While providing this support requires addressing a number of complex issues regarding the nature and feasibility of formative assessment, one issue that has been a particular focus of recent debate has been the domain-specificity of formative assessment. Can formative assessment be regarded as a more or less generic process, with professional development support for teachers being provided in a similar way to all teachers? Or, on the other hand, must support for teachers in their development of formative assessment practice be designed and delivered in a domain-specific way, with teachers of different subjects given different kinds of support?

The tendency in the research literature has been on the former approach, while the chapters in this book offer helpful illustrations of how formative assessment can be implemented in different domains. The authors of the contributions to this collection take different positions on the extent to which formative assessment needs to be treated as a generic process. Some, such as Heritage and Wylie (this volume), adopt a generic definition of formative assessment and explore how the generic processes can be implemented in a particular domain, while others, such as Jönsson and Eriksson argue that formative assessment needs to be defined in a way that is specific to the domain under consideration. In their chapter, Deans and Sparks (this volume), suggest that according generic aspects of formative assessment anything more than a minor role fundamentally undermines its potential, and that formative assessment must be discipline specific. In their different ways, each of the chapters provide a useful counterbalance to arguments that formative assessment is an essentially generic process

In this chapter, I want to argue for an intermediate position. I want to suggest that the best answer to the question of the domain-specificity of formative assessment is that formative assessment is irreducibly both domain-specific *and* domain-general, as implemented in the King's-Medway-Oxfordshire Formative Assessment Project (KMOFAP, Black, Harrison, Lee, Marshall, & Wiliam, 2003). In that project, a group of 24 (later 36) science and math teachers met every six weeks to explore formative assessment practice, with each one-day workshop including generic sessions, where teachers were introduced to central ideas about formative assessment, and domain-specific sessions, where teachers explored the implications of those generic principles for their subject. In other words, following Bennett (2011) I want to suggest that the answer to the question about the domain-specificity of formative assessment is "both/and" rather than "either/or."

I begin by briefly reviewing the development of formative assessment, and then discussing two central issues about how formative assessment should be defined. First, should formative assessment be defined descriptively or prescriptively: should definitions of formative assessment describe how the term is actually used, or should they provide guidance about how formative assessment should be implemented? Second, I discuss whether definitions of formative assessment should take into account what students should learn, how we decide what it means to know something, what happens when learning takes place or the instructional activities that students should engage in. The chapter concludes with a report of a large-scale randomized controlled trial of a generic approach to formative assessment that produced significant increases in student achievement at minimal cost (Speckesser et al., 2018), suggesting that the optimal approach to using formative assessment to improve student achievement at scale has to recognize that formative assessment is both domain-general and domain-specific.

Formative Assessment: Origins and Antecedents

Half a century ago, David Ausubel (1968) suggested that formative assessment was at the heart of effective instruction: "If I had to reduce all of educational psychology to just one principle, I would say this: The most important single factor influencing learning is what the learner already knows. Ascertain this and teach him [or her] accordingly" (Ausubel, 1968, p. vi). The same year, Benjamin Bloom outlined an approach to instruction that he called *mastery learning* (B. S. Bloom, 1968). The key idea in Bloom's approach was that perhaps as many as 90% of students could master what they were being taught in schools if their educational experiences were explicitly designed to achieve this goal. At the time, it was widely believed that high levels of educational achievement were possible only for the most able students, a belief that was no doubt in part reinforced by the high correlations observed between aptitude test scores and measures of educational achievement.

Rejecting such a belief, Bloom adopted John Carroll's definition of aptitude as the amount of time needed to attain mastery of a learning task under optimal conditions so that, given enough time, any student could attain mastery of a learning task, although Carroll himself acknowledged that the amount of extra time needed might be very great (Carroll, 1963). He estimated that a student at the 5^{th} percentile of aptitude would take five years to learn what would take a student at the 95^{th} percentile of aptitude just one year to learn.[1]

This represented a profound shift in perspective. With the old model, teachers taught, and some students learned the material and others did not, with the result that educational achievement was normally distributed. Under Bloom's mastery model, a normal distribution was a sign of instructional failure: "In fact, we may even insist that our educational efforts have been unsuccessful to the extent to which our distribution of achievement approximates the normal distribution" (B. S. Bloom, 1968, p. 3).

The reason that this shift was so profound, at least from the point of view of teaching, was that, with such a perspective, teaching became a *contingent* activity. However carefully teachers planned their instruction, they would not be able to predict in advance how much time would be needed by each of their students. Some form of assessment of the achievement of students would be needed to determine whether the required level of mastery had been reached. Assessment was now an integral part of instruction.

Of course the idea that students do not always learn what they are taught has probably been around for as long as people have been trying to teach others to do anything. And where instruction was individualized—as it was, for example in Frederic Burk's Individual System (Reiser, 1986), the Winnetka Plan (Washburne, 1941), the Dalton Plan (Parkhurst, 1922), or the Kent Mathematics Project (Banks, 1991)—the idea that the next steps in learning would be determined by the student's current level of achievement was an inherent feature. In such schemes, teaching was always a contingent activity. But such an approach was not common within the educational mainstream, and that is why Bloom's proposals were so radical.

In Bloom's approach, the time needed for a student to gain mastery would be revealed by periodic assessments of student progress:

> Much of what we have been discussing in the section on the effects of examinations has been concerned with what may be termed "summative evaluation." This is the evaluation which is used at the end of a course, term, or educational program. Although the procedures for such evaluation may have a profound effect on the learning and instruction, much of this effect may be in anticipation of the examination or as a short- or long-term consequence of the examination after it has been given.
>
> Quite in contrast is the use of "formative evaluation" to provide feedback and correctives at each stage in the teaching-learning process. By formative evaluation we mean evaluation by brief tests used by teachers and students as aids in the learning process. While such tests may be graded and used as part of the judging and classificatory function of evaluation, we see much more effective use of formative evaluation if it is separated from the grading process and used primarily as an aid to teaching. (B. S. Bloom, 1969, pp. 47-48)

The terms *formative* and *summative* had been proposed by Michael Scriven in response to a paper in which Lee Cronbach (1963) had suggested that asking an evaluator to determine the effectiveness of an educational program at the end of the development process was to offer the evaluator "a menial role and to make meager use of his services" (Cronbach, 1963, p. 3). Scriven (1963) pointed out that "there are many contexts in which calling in an evaluator to perform a final evaluation of the project or person is an act of proper recognition of responsibility to the person, product, or taxpayers" (Scriven, 1963, p. 7). Rejecting the idea that this was a menial role, he said, "It is obviously a great service if this kind of terminal evaluation (we might call it *summative* as opposed to *formative* evaluation) can demonstrate that an expensive textbook is not significantly better than the competition, or that it is enormously better than any competitor" (p. 5). In this context it is worth noting that this paper (or a revised version published some years later by the American Educational Research Association, cite) are often cited as the source of the term *formative evaluation* when, in fact, it seems that it was the term *summative* that Scriven was, in fact, proposing as novel.

It is also worth noting that while it is appropriate to attribute the formative-summative distinction to Scriven (in terms of curriculum, texts, or individual teachers) and Bloom (in

terms of students), it is important to realize that the qualifiers formative and summative did not represent new distinctions in the role that evaluation might play—in fact these ideas had been around for decades. What was new was the labels, as a way of clarifying debate.

Defining Formative Assessment

Since the pioneering work of Bloom, the idea that assessment can improve instruction as well as measure its results has become an important element in efforts to improve the educational achievement of students all around the world. Some authors (for example, Broadfoot et al., 1999) have suggested using the term *assessment for learning* in place of formative assessment, while others such as Calkins, Ehrenworth and Akhmedjanova (this volume) use the terms interchangeably. However, as Bennett (2011) has pointed out, this change merely shifts the definitional burden. More importantly, at least in the way the term is typically used, assessment for learning is a much broader term than formative assessment, as Black, Harrison, Lee, Marshall, and Wiliam (2004) explained:

> Assessment for learning is any assessment for which the first priority in its design and practice is to serve the purpose of promoting students' learning. It thus differs from assessment designed primarily to serve the purposes of accountability, or of ranking, or of certifying competence. An assessment activity can help learning if it provides information that teachers, and their students, can use as feedback in assessing themselves and one another, and in modifying the teaching and learning activities in which they are engaged. Such assessment becomes "formative assessment" when the evidence is actually used to adapt the teaching work to meet learning needs. (p. 10)

This is an important distinction because assessments that are given to motivate students (see, for example Assessment Reform Group, 2002) or to provide retrieval practice (see, for example, Roediger III & Butler, 2011) would, under many if not most definitions, be regarded as assessment for learning. However, it seems unlikely that many people would regard an assessment that was scheduled, but not administered, or administered, but not looked at further by either teachers or students, as formative.

Still others (Earl & Katz, 2006) have used the term assessment for learning more restrictively, to focus on the role of teachers, suggesting that the learner's role should be described as *assessment as learning* (Dann, 2002; Earl, 2003). While it is obviously attractive that students might be learning something while they are being assessed, as Bennett (this volume) points out, equating assessment with learning undermines the idea that both assessment for learning and assessment as learning should be regarded first and foremost as *assessment*. To equate assessment—probably best defined as a process of evidentiary reasoning (Mislevy, Almond, & Lukas, 2003)—with learning—defined by Kirschner, Sweller, and Clark (2006) as "a change in long-term memory" (p. 75) is unlikely to be helpful in clarifying debate.

More importantly, as Cizek, Andrade, and Bennett (this volume) note, not only is there no agreement about the terms that we should use, there is also no agreement about how to define formative assessment. This is not because people have not tried to define formative assessment. Indeed, there is no shortage of proposed definitions, and several of these are

discussed by Cizek et al. In my view, one of the most important reasons for the lack of agreement about the definition of formative assessment is because most of the definitions of formative assessment that have been proposed over the years, including those in the chapters of this volume, are, in essence, prescriptive. The desire for such prescriptive formulations is of course understandable, not least because it is natural to want to ensure that formative assessment is as effective as possible. Indeed, a major strength of the chapters in this volume, and a reason that they make important contributions to how we might improve formative assessment practice, is that they lay out in great detail how formative assessment might be implemented in particular disciplines, and these proposals seem to me to be eminently sensible. However, where such prescriptions are treated as definitions, the effect is to treat all approaches that do not conform to the definition as not being formative assessment, which is unfortunate for at least two reasons.

The first reason is that such an approach leaves little room for the creativity of teachers. For example, Cizek et al. acknowledge that "Whereas an assessment that was explicitly designed as a summative assessment could be used in a formative manner (and vice versa), that would clearly not be an optimal situation. Rather, a characteristic of any sound assessment is that it is used in the way it was designed" (p. xx). This may be good general advice, but as an empirical statement, it is unlikely to be correct.[2] For example, if a teacher were preparing a group of students for the College Board's Advanced Placement (AP) examination in History, the teacher might ask the students to take a practice test under formal test conditions. As well as providing an opportunity to familiarize the students with the test, such an occasion would also provide retrieval practice for the students, thus increasing their learning (Brown, Roediger III, & McDaniel, 2014). At the end of the allocated time, the teacher could collect and grade the students' responses, but an alternative would be to collect the responses and, several days later, give the completed response sheets back to the students and ask them, in groups of four, to compare their responses for each of the questions in the test and produce the best possible composite response. For students with incorrect answers, especially where they were confident their answers were correct, this exercise would lead to further learning via the hypercorrection effect (Butterfield & Metcalfe, 2001). Mindful of the work of Graham Nuthall (Nuthall, 2007) that showed that the advice given by a student's peers was often misleading at best, and often just incorrect, the teacher might then lead a whole class discussion in which each group shares its answer with the class and the teacher ensures that the students' understanding of the material is appropriate.

This use of the AP exam would appear to be one that is unlikely to have been envisaged by its designers, since the program was originally created to allow advanced high school students to earn college placement and course credit at higher education institutions. And yet, given that for many students the goal is to gain the highest score they can on the test, it seems perverse to regard this use of the AP test as inappropriate. Indeed, given the goals of the students, it might very well be optimal in terms of the additional learning generated and the time taken.

Another example of how a prescriptive definition leaves little room for the creativity of teachers, consider the definition of formative assessment proposed by the Council of Chief State School Officers (2008): "Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievements of intended instructional outcomes" (p. 3). While this

definition is attractive as guidance for how to implement effective formative assessment, any assessment that was used by a teacher to improve her instruction without the involvement of her students would not, under this definition, qualify as formative assessment. Clearly, for a number of reasons, the involvement of students in their own learning is desirable, but the CCSSO definition decrees that any use of assessment that is not shared by both teachers and students is not formative.

This problem is even more marked in the revised CCSSO definition adopted by Heritage and Wylie (this volume):

> Formative assessment is a planned, ongoing process used by all students and teachers during learning and teaching to elicit and use evidence of student learning to improve student understanding of intended disciplinary learning outcomes, and support students to become self-directed learners.
>
> Effective use of the formative assessment process requires students and teachers to integrate and embed the following practices in a collaborative and respectful classroom environment:
> - Clarifying learning goals and success criteria within a broader progression of learning;
> - Eliciting and analyzing evidence of student thinking;
> - Engaging in self-assessment and peer feedback;
> - Providing actionable feedback; and
> - Using evidence and feedback to move learning forward by adjusting learning strategies, goals or next instructional steps. (Council of Chief State School Officers, 2018)

While these are clearly helpful suggestions for improving the use of formative assessment in practice, the normative element here is particularly strong—and in my view, unhelpful. For example:

> if an assessment process is not planned, then according to this definition, it cannot be formative;
>
> if all students are not involved, it cannot be formative;
>
> if the assessment process helps teachers adjust instruction in a way that improves learning, but does not support students in becoming self-directed learners, then it is not formative;
>
> if the five listed practices are not integrated, then the process is not formative assessment.
>
> If the classroom environment is not collaborative and respectful, then there can be no formative assessment;
>
> if the learning goals are not located within a broader progression of learning, then there can be no formative assessment. In this context, it is worth noting that, as Calkins, Ehrenworth, and Akhmedjanova (this volume) point out, the learning progressions

> provided in the Common Core State Standards for opinion writing do not appear to reflect the developmental sequence followed by the majority of learners;
>
> if the assessment is used to improve non-disciplinary goals, such as well-being or mindfulness, then the assessment cannot be formative.

It is also worth noting that the second paragraph, in its use of the word "effective," makes a strong empirical claim for which no evidence is presented, and is unlikely, in fact, to be true, since the claim being made here is that where the five listed processes are not integrated and embedded, the formative assessment process will not be effective. In the KMOFAP project, participating teachers were explicitly required to focus their development efforts on just one or two strategies, and yet the project appears to have had a substantial impact on student achievement, improving student achievement on the English national school leaving examinations by around 0.32 standard deviations (Wiliam, Lee, Harrison, & Black, 2004).

Similar arguments can be made about the definition of formative assessment adopted by Calkins, Ehrenworth, and Akhmedjanova (this volume), which is adapted from Stiggins (2010), although it is worth noting that Stiggins proposed the definition as applying to all classroom assessment, including assessment for summative purposes, and not only formative assessment. Calkins et al. suggest that:

> formative assessment practices applied in classrooms should include: (1) clear purposes for assessing student work, (2) clear targets for what is being assessed, (3) high quality assessment designs for all assessment materials, (4) effective communication about what the results of assessment indicate about student learning, and (5) students' active participation in formative assessment. (p. X)

The implication here again is that assessment that does not, for example, require students' active participation cannot be formative.

To be clear, I am not objecting to any of the features of these different definitions as being ways to maximize the power of formative assessment. But to claim that these features *define* formative assessment is to render any practices that do not conform to a particular vision as not being formative assessment. Such a restrictive definition of formative assessment is at best unhelpful, and possibly completely counterproductive.

A second, and perhaps even more important, objection to the use of normative or prescriptive definitions is that they merely serve to perpetuate the definitional debate. As Cizek, Andrade, and Bennett (this volume) point out, several writers, such as Shepard (2008) and Popham (2006), have quite rightly criticized the use of the term formative assessment when used to claim legitimacy for something that is not supported by the research cited, as when research on classroom formative assessment is asserted to support the use of benchmark or interim tests as mechanisms for improving student achievement—a version of Kelley's jingle fallacy in which two things with the same name are assumed to be the same (Kelley, 1927). However, while the evidence often cited in support of benchmark or interim testing does not support the claims being made, it does not follow that the claims are not true—absence of evidence is not necessarily evidence of absence. In

this particular case, there are both logical and empirical arguments that show that benchmark or interim assessment can improve student achievement.

Start with the logical basis. It is obviously useful for school leaders to know whether students are actually learning anything in their schools, and the assurances of teachers that everything is on track are unlikely to be enough, so some kind of reasonably objective measure of student progress is essential to effective management of a school—after all any well-run organization should have ways of determining whether it is making progress towards its goals. And if different teachers of students in the same grade set their own assessments, it is difficult to compare results across classes, which is why the idea that teachers should create, curate, or adopt common assessments across their classes is so powerful. When some students are found to be making less progress than needed to be ready for the next grade, then appropriate action can be taken.

Moreover, there is empirical evidence that such approaches have been successful in improving student achievement. Saunders, Goldenberg, and Gallimore (2009) worked with instructional data teams in nine elementary schools in Southern California and found that, over a five-year period, students in these schools made significantly greater progress on standardized tests and other achievement measures than students in six comparable schools. They reported an effect size of 0.79 at the teacher level. Assuming a correlation of 0.15 between teacher quality and student achievement (Hanushek & Rivkin, 2010), this would represent an increase in student achievement of 0.12 standard deviations. Using norms developed by H. S. Bloom, Hill, Black, and Lipsey (2008) for children in third through fifth grade, such an effect size would represent an increase in the rate of learning somewhere between 20% and 40%. Given the time invested by the teachers, such an intervention could be a highly cost-effective way of improving student achievement. Similar benefits of the use of interim or benchmark assessments were found by Barry and Leslie Pulliam in the Focus on Standards program, which involved teachers comparing student performance on tests with their state standards, and their curriculum resources (Goe & Bridgeman, 2006). So while Shepard and Popham are quite correct to point out that the use of benchmark and interim assessments is not justified by the research that is generally cited, it is not correct to say that such uses have no evidence in their support.

In response, it could be argued that the term formative assessment should be reserved for uses of assessment that are relatively close to instruction, and a term such as formative evaluation could be used to describe more distal uses of assessment to improve learning. However, given that those advocating for interim and benchmark assessment have substantial investments in the term “formative assessment” (DuFour, 2007), it seems unlikely that calls for such definitional clarity will be heeded.

To summarize, while it is clearly a matter of judgment whether formative assessment is defined descriptively or prescriptively, defining formative assessment prescriptively is likely to result in the exclusion of many uses of assessment that do, in fact, improve learning, and which many people would regard as formative. Whatever one thinks about the use of benchmark or interim tests to monitor student progress and to align curriculum, such assessments are, in the literal sense of the term, functioning formatively, especially when such tests are keyed to the instruction that students have received as the were in the work of the Pulliams. Evidence of achievement is being elicited, interpreted, and used to make decisions about instruction that are likely to benefit students. To say that such assessment

processes are not formative is to redefine the term *formative* in a way that is completely at variance with its use in natural language, and indeed, at variance with the sense that the term was proposed by Scriven and Bloom.

A prescriptive definition of formative assessment will therefore, in my view, make it harder for academics and practitioners ever to reach an agreed definition. It would, of course, be wonderful if—upon being told that what they regard as formative assessment is not, in fact, formative assessment—those individuals and entities with different views accepted this, stopped using the term, and looked for another term to describe their practices. It seems to me, however, that this is extremely unlikely. Rather, restricting the definition of formative assessment to a subset of the ways in which it is currently used practically guarantees that no agreed definition will ever be established. Instead, what is needed is to develop an inclusive definition of formative assessment that excludes none of the processes that are, within reason, described as formative assessment. Put bluntly, we should not make the word *formative* work too hard. Formative should just mean formative, so that we can then focus on the features that make formative assessment more or less effective.

As a matter of practical necessity, therefore, formative assessment needs to be defined in a descriptive and inclusive way, rather than in a prescriptive, normative manner. The next issue that arises is the scope of the definition. As each of the chapters in this volume have pointed out, formative assessment practices must take account of the domain being assessed. What is less clear, however, is whether the nature of the domain entails any commitments about how formative assessment should be defined.

Four Issues about How We Define Formative Assessment

The next section addresses four questions about how formative assessment can or should be defined, and in particular examines whether commitments to formative assessment necessarily entail a particular view of what students should be learning, what it means to know something, what happens when learning takes place, and what kinds of pedagogical activities teachers should arrange for their students.

Does a commitment to formative assessment entail a commitment about what is to be learned? Most of the chapters in this collection discuss aspects of formative assessment from a particular set of assumptions about what is to be learned. Bennett (this volume) suggests that "reasoning about assessment design starts with articulating the claims to be made from assessment results about individuals or institutions" and that "those claims should derive directly from state content standards, cognitive-domain theory, curriculum frameworks, learning objectives, or some combination of these sources" (pp. 2-3).

While such clarity about educational outcomes may well be desirable, as Calkins et al. (this volume) and Andrade et al. (this volume) point out, sometimes it is not possible to codify the quality of work in formal standards, theories, curriculum frameworks, or learning objectives. Sometimes, the best that we can do is to help students develop what Claxton (1995) describes as a "nose" for quality. Students become enculturated into a particular community of practice in which their teachers are already participants. Effective formative assessment then requires teachers to possess, in addition to a shared construct of quality, an understanding of the "anatomy" of quality, so that they can identify instructional next steps

that may not be related in any obvious way to the goal, but do help students to progress (see discussion of cognitive load theory below).

Moreover, in many parts of the world, teachers determine their instructional goals not by reference to formal standards or curriculum frameworks of learning objectives, but by reference to the examinations their students need to pass. In many countries, these examinations are accompanied by examination syllabuses that specify what may, and may not, be assessed, but typically these are written at a level of generality that provides little guidance for teachers. For this reason, teachers determine priorities for instruction by looking at the examination papers that have been set in previous years (which are usually readily available and accessible). Even in countries like the US, where state content standards are provided, the relationship between what is specified and what is assessed is far from straightforward. Sometimes this ambiguity is due to the technical inadequacy of the assessments used in that state, but often it arises from a mismatch between the specified content and the means of assessment. For example, Common Core state standard number 8(c) for mathematics requires students to "design and use a simulation to generate frequencies for compound events" (Common Core State Standards Initiative, 2010, p. 51) . However, whether this could be meaningfully assessed in a standardized test, especially one that relied on multiple-choice items, is doubtful. Where teachers are under pressure to raise test scores, especially given the fact that most state content standards contain more content than most students can learn in a year, teachers may well choose not to spend time on material like this that is unlikely to be assessed. This is not to condone such behavior, but merely to point out that what a particular teacher is trying to achieve cannot be determined by the officially mandated curricula that are in place.

The two chapters in this collection that address science education (Jönsson and Eriksson, this volume; Furtak, Heredia and Morrison, this volume) define the goals of science education so as to include knowledge about science as a social and cultural practice, as well as the concepts, theories and models that scientists have generated. In particular, Furtak et al. show that, around the world, there is increasing consensus that science education should include both the things that scientists have found out, and how such knowledge is generated. However, it is important to note that these are arbitrary choices (in the original sense of requiring judgment). In many countries, science education does not include science as a cultural and social practice. Any approach to formative assessment that entails an acceptance of a particular definition of what students should learn is likely to be unhelpful to those who define science, or for whom science is defined, in a different way.

In a similar vein, Jönsson and Eriksson (this volume) draw three distinctions between higher education and earlier phases of education, in terms of the autonomy of the learner, the size of classes, and connections to research. First, while it is certainly true that many of those involved in teaching in higher education see the development of learner autonomy as a key outcome, others do not, and many elementary school teachers would also claim that learner autonomy is a key aim for them and their students too. The *extent* to which autonomy is a goal for education may vary from one phase of education to another, but it is a difference in degree, rather than kind. Second, while classes in undergraduate education can be very large, they are often not, with seminar groups often composed of twenty or fewer students. And in sub-Saharan Africa, classes of 70 or more are common in elementary schools. Again, the variation within each phase is much greater than the difference between the phases. Third, while it is true that many higher education

institutions do claim that having students being taught by active researchers is a benefit, there is little evidence to show this belief is true, and, in most higher education institutions in the US, let alone the rest of the world, research is a marginal activity for those involved in teaching undergraduates (and teaching is often a marginal activity for those involved in research).

As a third example, Andrade et al. (this volume) define arts education as focusing on creativity and while such a view may well be agreed to by many, and perhaps almost all, arts teachers, it is far from clear whether this stipulation is essential. For example, many instrumental music teachers do not regard creativity as particularly important, at least while a student is learning an instrument.

If we are to have broadly agreed definitions of formative assessment, it therefore seems essential that a commitment to formative assessment entails nothing about what students are to be learning, since different teachers may well have different goals. State standards include far more material for each grade than most students can learn—presumably because the standards are designed to keep the fastest-learning students occupied for the whole year. But, following Carroll (1963), this position means that there is far too much for most students to learn, and so teachers have to make choices about what to teach. Such choices will depend on the kinds of state tests in place, and even such prosaic factors as whether the teacher has tenure. Teachers are placed in an impossible position, and have to make compromises, taking a number of factors into account, and so teachers do need to be clear about what their students are to learn—in fact, formative assessment can only begin once the teacher is clear about the purpose of instruction. But, as noted above, any approach to defining formative assessment that entails a particular set of assumptions about what students are to learn renders formative assessment irrelevant in many settings. Given the reasonably clear evidence about the effectiveness of formative assessment to raise student achievement even where such achievement is measured though standardized tests, anything that dissuades teachers from embracing formative assessment because of the restrictions applied is, in effect, a way of lowering student achievement.

Does a commitment to formative assessment entail a commitment about what it means to know something? As well as making assumptions about what students should be learning, most of the chapters in this volume also make a number of epistemological assumptions—assumptions about what it means to know or understand something. Most science curricula require that, at some point in their school careers, students learn Archimedes' principle. However, what is often less clear is what it means to know Archimedes' principle. At one level, we might be content that a student can recite the principle in the standard form:

> "Any object, wholly or partially immersed in a fluid, is buoyed up by a force equal to the weight of the fluid displaced by the object."

A student who can recite this definition could be said to "know" Archimedes' principle, but others, such as Gobert et al. (2011) have argued that while knowing science does require knowing how to state scientific principles, students also have to be able to use the principles to reason scientifically. Even here, quite what we mean by being able to use a principle varies. We might say that a student knows Archimedes' principle if she can use it

to explain why ice floats, or we could be more demanding and say that they only really know it if they can use it to answer questions such as the following:

> Someone sits in a boat in a swimming pool holding a 10kg mass. What happens to the level of the water in the pool if the mass is dropped into the swimming pool?

Being able to answer such questions would entail a far deeper knowledge of the subject matter than just being able to answer simple questions about why ice floats, although the basic concepts are identical.

The important point here is that people may reasonably disagree about what it means to understand Archimedes' principle, and so, in that same way that a commitment to formative assessment cannot entail any particular commitment about what students are to learn, a commitment to formative assessment cannot entail any commitment about what it means to know. Those ideas must be clarified before formative assessment can begin.

Does a commitment to formative assessment entail any view about what happens when learning takes place? In recent years, a number of authors have suggested that a commitment to formative assessment necessitates a socio-cognitive, or a socio-cultural perspective on psychology (see, for example, Shepard, Penuel, & Pellegrino, 2018), but again, whether such commitments are necessarily entailed is open to question, and depends on one's views about what happens when learning takes place.

In the 1960s, the prevalent idea that learning was simply making links between stimuli and responses came into question because such a model would predict that students' errors should be random, whereas, particularly in mathematics and science education, students' errors were to a significant extent predictable (Driver & Easley, 1978; Hart, 1981). Students were not "misremembering" what they had been taught but were rather constructing their own knowledge on the basis of their experience of the world. Learning was an active "constructive" process. Constructivism, as a view of what happens what learning takes place, accounted for many observed phenomena, such as misconceptions, that were not explained well by associationist views of learning. However, constructivism was not particularly effective in explaining aspects of learning, such as learning number facts, that associationist approaches explained well. In psychology, each new theory about what happens when learning takes place is very good at explaining things that the pre-existing theories did not, but are often not very good at explaining what the pre-existing theories explained well. Unlike in science—where newer theories tend to subsume previous theories, in the way that Einstein's approach to physics includes Newton's ideas as special cases—in psychology, new theories tend to provide different perspectives rather than more complete solutions. This is why Anna Sfard (1998) stresses that we need multiple—often incommensurable—perspectives on learning, rather than trying to figure out which is the best, or more complete theory.

For example, when students learn that there is a single electron in the outer shell of a sodium atom, which is why it bonds with a single atom of chlorine to make salt, there is not much to "understand" here. There is no point in asking why an atom of sodium has a single electron in its outer shell. Students just need to know that the atoms of the element that we call sodium happen to have a single electron in the outer shell. If, on the other hand, we are

trying to figure out why many young children's believe that wind is caused by the movement of trees, then associationist approaches are unhelpful. This "misconception" is not the result of poor-quality science instruction, nor is it the result of insufficient reinforcement of the correct links between stimuli and responses. Rather it is the result of students creating schemas to make sense of their experience. While some people adopt entrenched positions on such issues, it probably makes more sense to acknowledge that some aspects of learning science are best described by the former approach, which we might call an associationist approach, while some are more like the latter which we might call a constructivist approach.

A commitment to formative assessment, therefore, should not entail any particular commitment to what happens when learning takes place. Formative assessment is necessary, because, as Ausubel (1968) noted, good instruction starts from where the learner is, and because, what students learn as a result of any particular sequence of instruction is impossible to predict with any certainty. For those who believe that learning is a matter of making associations between stimuli and responses, then it is impossible to predict in advance how much reinforcement will be required before the associations are established, so establishing what has been learned and then taking appropriate remedial action is essential. For those who believe that students construct their own knowledge about the world, then we need to find out what sense the students have made of their instructional experiences. For those who adopt situated perspectives on learning, it is necessary to determine the extent to which a student's performance in a particular task is the result of attunements to affordances or constraints in a particular environment. Those with different views about what happens when learning takes place may have different reasons for using formative assessment, but a commitment to formative assessment cannot entail a commitment about what happens when learning takes place.

Does a commitment to formative assessment entail any commitment about how students should be taught? Several of the chapters make implicit claims about the kinds of pedagogical activities that are likely to promote strong disciplinary learning. For example, Burkhardt and Schoenfeld (this volume) suggest that widely employed imitative methods such as "I do, we do, you do" "cannot address the now-generally-accepted importance of extended reasoning, non-routine thinking, and problem solving" (p. 2). No evidence is provided in support of this claim, and recent work in cognitive psychology suggests that it may be incorrect. For example, John Sweller and his colleagues have provided considerable evidence that, for novices, worked examples, with guidance being progressively faded, result in superior learning about mathematical problem solving than more "authentic" activities (Sweller, Kalyuga, & Ayres, 2011). While the full implications of cognitive load theory are still being explored, there is now considerable evidence that in many cases, inquiry-based approaches to learning may be less effective, at least for some learners.

Similar arguments apply in the context of attempts to make science more interesting to students through the use of more inquiry-based approaches to instruction, and especially those that attempt to engage students with "real" contexts, such as the red fox task discussed by Furtak et al. (this volume). The task was developed by a group of high school biology teachers "to connect with students' lived experiences in observing red foxes in the communities around their homes" (p. 7). Where students are not particularly interested in

science, then approaches of this sort may be necessary to increase student engagement, but it is important to note that while inquiry-based instruction appears to increase student engagement in science, its impact on student achievement is far from clear. Indeed, in the most recent round of the Programme for International Student Assessment (PISA), the prevalence of inquiry-based instruction was negatively correlated with student achievement in science (Organisation for Economic Co-operation and Development, 2016: Figure II.7.2). Now of course it could be that teachers who teach low-achieving students make greater use of inquiry-based instruction in order to increase student engagement—the direction of causality is unclear—but it does at least raise questions about what makes for effective instruction.

My aim here is not to present arguments for or against the respective merits of different approaches to instruction. Rather it is to point out that, given the great uncertainty about what kinds of instructional approaches work best, and in which circumstances, a definition of formative assessment that entails any commitment about what kinds of learning activities should be employed immediately condemns itself to irrelevance in many settings.

The rather extended discussion of these four questions leads, inevitably, I believe, to the conclusion that any definition of formative assessment should entail absolutely no commitments whatsoever about what students are to learn, what it means to know, what happens when learning takes place, and what activities are likely to be most effective in getting students to learn. A commitment to formative assessment does, to be sure, require that we take into account what students are to learn, what it means to know, what happens when learning takes place, and what kinds of instructional activities are likely to be successful. But we should not define formative assessment in a way that, for example, rejects as illegitimate a belief that history is about learning facts and dates, that knowing simply means being able to reproduce what one has been taught, that learning is just making links between stimuli and responses, and that students learn best through lectures. For the avoidance of doubt, these are not positions that I am defending—in fact I disagree with each of these propositions. What I am saying is that debates about the adequacy of such ideas should be discussed separately, and should not be smuggled in through the way we define formative assessment.

If we are to maximize the power of formative assessment to improve student learning, we need to frame definitions that include, rather than exclude, and encompass all the ways that the term formative assessment is used. It was this concern to be inclusive that led Paul Black and myself (2009) to suggest that assessment functions formatively "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" (Black & Wiliam, 2009, p. 9).

A full description of the rationale for this definition can be found in Black and Wiliam (2009), but for the purpose of the present discussion, the important feature of this definition is that it is inclusive of all the ways that the term formative assessment is currently used, from interim or benchmark assessments administered monthly or even quarterly and used to align curriculum and monitor student progress, to the minute-by-minute use of assessment to judge the level of understanding of a particular concept in a group of students. It focuses on assessment as measurement, as advocated by Bennett (this volume) and its role in

evidentiary reasoning, and does not require that the assessment actually improves learning—only that it is likely to. This definition is also consistent with the idea that the formative-summative distinction is a distinction about the kinds of inferences supported by assessments, rather than in terms of assessment instruments, or assessment outcomes (Black & Wiliam, 2018).

Adopting an inclusive definition matters, because discussion can then move on from the relatively unproductive boundary disputes about whether certain practices are, or are not examples of formative assessment to the more important and substantive issues about whether, and if so, by how much, and under what circumstances, student achievement is increased through the use of such assessment. In particular, given the focus of this book, it is important to discuss the extent to which formative assessment can be successfully implemented as a domain-general process and the extent to which student learning is improved by emphasizing the idea of formative assessment as a domain-specific process.

Before further discussing the extent to which it is helpful to regard formative assessment as domain-specific versus domain-general, it is important, to point out that the terms *domain*-specific and *discipline*-specific are not equivalent. Good formative assessment in science looks different from good formative assessment in English language arts at least partly because of the differences in the subject matter being taught. However, it is important to note that while some of those differences are inherent in the discipline, others are not. For example, in the teaching of science in most countries, more time is spent teaching students about the knowledge that science has generated than on how science generated that knowledge. It is therefore not surprising that formative assessment practices in science classrooms often bear strong similarities to formative assessment practices in math classrooms, with, for example, a focus on alternative conceptions or facets of knowledge (Minstrell, 1992). In contrast, particularly in the Anglophone countries, English language arts instruction focuses on students' experiences, and their personal responses to text, so that formative assessment practices often look very different from those in math and science classrooms. However, it is important to recognize that the ways that school curricula are developed represent a series of choices about what students should be learning in school—in Denis Lawton's memorable phrase, a curriculum is "essentially a selection from the culture of a society" (Lawton, 1975, p. 7)—and that different choices could have been made. For example, in the second half of the nineteenth century, it was common to find students in English language arts classrooms analyzing sentences by drawing diagrams (Edgar, 1915). The kinds of formative assessment practices that would be used in such a classroom would resemble much more closely those used today in science and math classrooms than those used in English language arts classrooms. In contrast, if a class were discussing the ethical implications of, say, genetically-modified foods, the discussion would resemble the kind of discussions common in English language arts and social studies classrooms, with consequent implications for the most appropriate formative assessment practices.

The important point here is that different ways of defining a school subject will result in different implications for formative assessment practice. Knowing which *discipline* is being taught tells us very little about what kinds of formative assessment practice would be appropriate. Knowing how the discipline has been defined within the school curriculum, on the other hand, would be much more informative. The *domain* matters much more than the *discipline*.

Domain-General and Domain-Specific Approaches to Professional Development for Formative Assessment: Empirical Evidence

In debating the respective merits of domain-general and domain-specific approaches to formative assessment, it is important to note that the extreme positions in this debate—that formative assessment is entirely domain-specific or entirely domain-general—are demonstrably absurd. Formative assessment is, obviously, both domain-general and domain-specific. The idea that a teacher's instructional decisions are likely to be better if evidence about current understandings are collected from all the students in a group, rather than just the individuals who are confident enough to volunteer answers can be applied to any group teaching situation, and so as a technique is completely generic. I do not need to know what you are teaching to know that evidence from all the students in the group about their current level of understanding of the material being taught is likely to result in better instructional decisions. At the other extreme, the questions that teachers need to ask to determine whether their students have understood the material being taught can only be determined with substantial content knowledge. As has been said several times, what is important, therefore, are the trade-offs that occur when we treat formative assessment as more generic or more subject specific.

The chapters in this volume clearly demonstrate the advantages, for teachers and for students, of operationalizing formative assessment in a domain-specific way. However, there are also a number of disadvantages. First, it makes it more difficult for teachers to learn from teachers of other subjects, since a common language of description is less likely to be developed (Black et al., 2003). Second, students' experiences in school will become less coherent, since teachers in different domains may define the same processes differently, and different processes may be defined similarly (Kelley's "jingle-jangle" fallacies again). Third, administrators are likely to find it more difficult to support teachers, and getting access to the domain-specific expertise is likely to be challenging, particularly in small school districts, and for domains outside the core subjects of math, English language arts, science and social studies. Fourth, developing whole-school policies is likely to be more difficult, making system-wide implementation of formative assessment more difficult to secure (Thompson & Wiliam, 2008).

The challenge, then, is to define formative assessment as a generic process, but to do it in a way that accommodates domain-specific definitions and practices that are, as far as possible, both consistent with the generic approach, and do not conflict with definitions and practices that are used in other domains. Or, to put it another way, we need to determine the extent to which formative assessment can be usefully and productively defined and operationalized as a generic process, and to what extent it is necessary for the practice of formative assessment to take into account the discipline or the domain.

One way to achieve such a framework would be to survey practice in different disciplines and look for commonalities. As Jönsson and Eriksson (this volume) point out, many approaches to formative assessment in effect do exactly that. The result is that any

collection of practices thus identified will lack a strong theoretical foundation which makes it more difficult to identify whether different techniques conflict, and it is also impossible to determine whether the totality of practices is complete.

To provide a theoretically-grounded approach to the operationalization of formative assessment, I and my colleagues began by identifying what we thought would be more-or-less universally agreed assumptions.

We began by looking at formative assessment as a process of closing the gap between current and desired levels of achievement as suggested by Ramaprasad (1983) and D. R. Sadler (1989). However, while this idea seemed to be acceptable to many teachers, some teachers, particularly those from English language arts, and the creative arts (art, music, dance, drama) found the idea of a "gap" between current and desired levels of achievement an unhelpful way of thinking about their practice (Marshall, 2000).

For this reason, when we sought to theorize formative assessment, we looked for starting points that would be likely to be agreed by all teachers. As an absolute minimum, it seemed to us that teaching should be an intentional process. In other words, the teacher should have some idea of what kinds of changes they seek to effect in their students. These ideas might be expressed as learning targets, goals, aims, or objectives, but could also take the form of implicit understandings of what it means to participate in a particular domain (Sfard, 1998). Whether or not they can be expressed in words, any instruction should start with some intentions about what students should learn.

Once the intentions have been defined, then, returning to Ausubel's (1968) principle for instruction, the crucial processes are to establish:

1) Where the learners are in their learning
2) Where they are going
3) How to get there

Crossing these three processes with the three kinds of agents in the classroom—the student, their peers, and the teacher—produces a 3 x 3 grid, which can be simplified to yield five strategies of formative assessment, as shown in Figure 1 (Leahy, Lyon, Thompson, & Wiliam, 2005; Wiliam & Thompson, 2008).

<table>
<tr><td></td><td>Where learner is going</td><td>Where learner is now</td><td>How to get the learner there</td></tr>
<tr><td>Teacher</td><td rowspan="2">Clarifying, sharing, and understanding learning intentions and criteria for</td><td>Eliciting evidence</td><td>Providing feedback that moves learning forward</td></tr>
<tr><td>Peer</td><td colspan="2">Activating students as learning resources for one another</td></tr>
</table>

Learner	success	Activating students as owners of their own learning

Figure 1: Five strategies of formative assessment (Wiliam & Thompson, 2008)

This framework provides a useful way of integrating research literature from a number of fields (Education Endowment Foundation, 2018), and, most important, seems to be accessible to teachers, and useful in their daily work. As just one example, this framework has been adopted by the Singapore Ministry of Education as a central element of its Primary Education Review and Implementation—Holistic Assessment (PERI-HA) program, and has been implemented in over two-thirds of the nation's elementary schools (Tan, Teng, Tan, & Peng, 2014).

In addition, while some of the original nine cells have been merged in order to make the framework easier to use and apply, as a study by Chen, Lui, Andrade, Valle, and Mir (2017) discussed by Andrade et al (this volume) shows, the merged cells can be unmerged in order to make important distinctions in particular contexts. However, perhaps the main strength of this framework is as a general one for thinking about formative assessment that applies to all disciplines, allowing the specificities of each discipline or domain to be honored within a coherent structure.

Ultimately, however, the argument about the extent to which formative assessment needs to be conceptualized in a domain-specific way, or can be treated generically, is, at its heart, an empirical question. Most of the chapters in this collection argue that student learning is enhanced when formative assessment is conceptualized in a discipline-specific, or domain-specific way. Now at one level, as mentioned earlier, this is obviously true. Formative assessment has to be conceptualized in a domain-specific way. Formative assessment involves the collection of evidence about student achievement, and without a clear idea about what students should be learning, it is impossible to know what evidence would be relevant. And while Furtak et al. (this volume) note that many science educators have objected to the term "misconception" for describing students' non-standard or incomplete ideas about scientific topics, students do seem to learn more when they are taught by teachers who know the misconceptions that students are likely to have (P. M. Sadler, Sonnert, Coyle, Cook-Smith, & Miller, 2013). Similarly, while it is possible to formulate general principles about good feedback such as "make feedback into detective work" (Wiliam & Leahy, 2015, p. 124), actually doing this requires detailed subject knowledge, as does activating students as learning resources for one another, and as owners of their own learning. This is what prompted Paul Black, myself, and our colleagues to co-author a series of short booklets for teachers with subject-specialists in English (Marshall & Wiliam, 2006), mathematics (Hodgen & Wiliam, 2006), science (Black & Harrison, 2002), modern foreign languages (Jones & Wiliam, 2007), geography (Lambert & Weeden, 2006), information and communications technology (Cox & Webb, 2007), design and technology (Moreland, Jones, & Barlex, 2008) as well as a similar booklet for elementary school teachers (Harrison & Howard, 2009).

But the arguments about the necessity for domain-specific approaches to formative

assessment made in this collection are, implicitly, making a stronger claim. They are arguing that it is necessary to draw out the implications of generic strategies for particular domains in order to help teachers implement formative assessment more effectively. After all, if defining formative assessment in a subject-specific way, or drawing out the domain-specific implications of particular strategies, had no impact on student achievement, there would be little point in doing so.

The key question, therefore, is to what extent providing guidance to practitioners about the implementation of formative assessment practices in specific domains increases the impact on student achievement, and this is the focus of the remainder of this chapter .

First, it is worth pointing out while the chapters in this collection present thoughtful ways of implementing formative assessment in the disciplines that are consistent with the research evidence, the majority of the chapters present little or no empirical evidence that domain-specific approaches to the development of formative assessment are in fact more effective.

As Burkhardt and Schoenfeld (this volume) report, the 8 to 12 days devoted to "formative assessment lessons" developed by the Mathematics Assessment Project resulted in increased student achievement, but it is not clear whether this can be attributed to formative assessment, since the intervention included, in addition to encouraging and supporting the use of formative assessment, a series of high-quality lesson plans, which may have made a substantial contribution to the improvements in achievement (Jackson & Makarin, 2016). The project evaluators (Herman et al., 2015) found that the mathematics achievement of students working with the formative assessment lessons was 0.13 standard deviations higher than comparable students, which they equate to an extra 4.6 months learning into months of learning, using norms derived by Howard Bloom and his colleagues from standardized tests (H. S. Bloom et al., 2008; Hill, Bloom, Black, & Lipsey, 2007). However, the tests used to evaluate the formative assessment lessons were developed by the Mathematics Development Collaborative and, because of the way the tests were developed, they are likely to be more sensitive to the effects of instruction than the traditional standardized tests analyzed by Bloom and his colleagues. As a result the effect of the formative assessment lessons may be somewhat overestimated.

More positively, Andrade et al. (this volume) report on studies undertaken through the Arts Achieve project that showed students taught by teachers receiving a two-year professional development program focused on formative assessment made more progress than students taught by other teachers. Although the study was designed as a cluster-randomized trial, the low fidelity of implementation by some teachers led the researchers to use propensity-score matching to compare teachers who implemented the program with fidelity to matched teachers who were not exposed to the program, and found a net effect of 0.26 standard deviations. Interpreting this result is not straightforward for several reasons. First, the teachers who did not implement the program with fidelity may be less competent, and so it is not possible to attribute the effects to the professional development. Second, since the project included elementary, middle, and high school teachers, it is not clear what kinds of norms would be most appropriate to convert the effect size into a rate measure such as the

number of extra months of learning. Third, the assessments used in the study may have been, like the measures used by Burkhardt and Schonfeld, more sensitive to the effects of instruction. Nevertheless, this result represents a substantial increase in the rate of learning, and close to the median found for formative assessment programs in other subjects (Kingston & Nash, 2011, 2015).

To examine the empirical claim that teacher professional development in formative assessment needs to be domain-specific, it is instructive to compare the effect sizes reported by Burkhardt and Schoenfeld and by Andrade et al. with those obtained from a recently conducted evaluation of a generic, whole-school professional development program.

Embedding Formative Assessment (EFA, Leahy & Wiliam, 2013) is a two-year professional development program developed from the Keeping Learning on Track program discussed by Andrade et al. (this volume). The EFA program was designed to be delivered within schools with minimal additional cost, and without any external facilitation, not because such facilitation would not be helpful, but rather because, in many local education authorities, particularly in the United States, such additional resources cannot easily be found from school budgets. Even where such funds are available at a particular point in time, budgets are often volatile, and where cuts have to be made quickly, professional development appears to be a particularly convenient place to make them. For this reason, it was determined that to be sustainable, a teacher professional development program could not require anything more than minimal additional cost.

The EFA program consists of all the materials and handouts needed to run 18 monthly Teacher Learning Community (TLC) meetings of 75 to 90 minutes duration, over a two-year period, together with videos of classroom practice exemplifying formative assessment practices, interviews with educators, administrators and students, and a variety of other relevant resources.

Through the EFA program, teachers are introduced to the five strategies of formative assessment suggested by Leahy et al. (2005) and shown in Figure 1 above. At each meeting, each TLC member commits to trying out at least one strategy in their classroom, and at the next meeting, the following month, they report back to their peers on their experiences. Teachers are also encouraged to observe each other and give each other feedback on their development of formative assessment.

While the program has been implemented in a number of countries, including Sweden, Australia, England and Scotland, and was found to be helpful by teachers and administrators, there was little more than anecdotal evidence that the program actually increased student achievement. Accordingly, the Education Endowment Foundation—a UK-based philanthropic organization—funded an evaluation of the program in secondary schools in England, which was awarded to the UK's National Institute for Economic and Social Research (NIESR).

The NIESR estimated that a cluster-randomized trial of the program with an 80% chance to detect an effect size of 0.2 standard deviations would require 140 schools, since randomization would need to take place at the school level (as teachers in the same school

could be expected to talk to each other). The study took the form of a preregistered "intention to treat" study[3], with half the schools allocated at random to receive the EFA materials, and the other half being given the cash equivalent of the cost of the materials (approximately $500 at the exchange rate prevailing at the start of the study). This last feature is particularly important, because the analysis took no account of the fidelity of implementation in the experimental schools. The performance of all students in the study cohort in the schools allocated to the experimental group was compared to that of all students in the control group, unlike the study reported in Andrade et al. (this volume).

After schools had been recruited, and allocated to either the treatment or control group, it was discovered that some of the schools had already participated in a professional development program titled the Teaching Excellence Enhancement Programme (TEEP), which included many elements of the EFA program. It was decided that these schools should not be included in the main analysis, leaving a total of 58 schools assigned to receive the EFA material, and 66 schools receiving the cash equivalent.

The measure of achievement used in the study was the average grade (on a nine-point scale) received by students in England's national school leaving examination, the General Certificate of Secondary Education (GCSE), in mathematics, English, and their best six other subjects[4]. This composite measure, which is called Attainment 8, is the key measure that is used to hold schools accountable for the academic achievement of their students and the grades that students achieve on their GCSE examinations are also important determinants of their options for future study and employment. The outcome measure that was used in this evaluation is therefore a measure that is of great concern both to schools and to their students. In total data were collected on 22,709 students in the participating schools who commenced their studies in September 2015 and took their school leaving examinations in June 2017.

A full description of the research protocols can be found in Speckesser et al. (2018). The primary analysis consisted of fitting two models to the data:

a "simple" model, including prior attainment and allocation dummy variables as fixed covariates, with school as a random effect.

a "precise" model, including prior attainment, the allocation dummy and indicator variables specifying membership of the randomization blocks (all fixed effects) and schools as a random effect

The average Attainment 8 scores for students in the original 70 treatment schools were 0.10 standard deviations higher than for those in the original 70 control schools, and this result was not statistically significant. However, when comparing the 58 experimental schools and the 66 control schools that had not been exposed to the TEEP program—arguably a fairer comparison of the effect of the EFA program—students in the experimental schools scored 0.13 standard deviations higher, and the result was statistically significant (p=0.04).

To interpret this effect in terms of increases in the rate of learning, it is necessary to estimate the progress made over two years by students in the control group. These students were 15 years old at the beginning of the trial. The norms produced by H. S. Bloom et al. (2008) discussed above suggest that one year's progress for 15 year old students is

approximately 0.2 standard deviations, while NAEP scores increase by about one standard deviation over four years, suggesting an annual equivalent of 0.25 standard deviations (National Assessment of Educational Progress, 2013). On the other hand, Rodriguez (2004) estimates that for eighth grade students on the mathematics tests used in the Trends in Mathematics and Science Study, one year's progress is 0.36 standard deviations. Finally, the Organisation for Economic Co-operation and Develepment assumes that one year's growth for 15-year-olds on the tests used in the Programme for International Student Assessment (PISA) is 0.3 standard deviations (Andreas Schleicher, personal communication, November 14th, 2018). Given these results, it seems reasonable to assume that one year's progress for students in the control group would be in the range of 0.2 to 0.4 standard deviations, with 0.3 as a reasonable central estimate.

Since the program spanned two school years, it is necessary to make some allowance for the attrition of student learning from the first to the second year of students' GCSE studies. In a meta-analysis of 39 studies, mostly from the 1970s and 1980s, Cooper, Nye, Charlton, Lindsay, and Greathouse (1996) found an attrition of around 10% of learning from one year to the next. However, other studies have found much larger estimates. For example, using data from the North West Evaluation Association's Measures of Academic Progress test, Thum and Hauser (2015) found attrition rates of 25% for reading and as much as 40% for math. Given this, it would appear that assuming an attrition rate of 10% is conservative.

Students in England take their GCSE examinations half-way through the third and final term of the English academic year—in other words, the final year of the GCSE program is really only five-sixths of a year, so over the two years of their GCSE studies, students in the control group could be expected to increase their achievement by

first year growth x 0.9[to account for attrition] + second year growth x 5/6 [to account for the shorter second year]

Using the range of estimates of annual growth from 0.2 to 0.4 discussed above yields an expected increase in the range of 0.35 to 0.69, with a central value of 0.52. Since the effect of the EFA program over the two years was to increase student achievement by 0.13 standard deviations, this suggest that the program increased the rate of student learning between 19% and 38%, with a central estimate of 25%. Given that this increase is the average across all students in the experimental group—not just those who implemented the program with fidelity—this is an important finding. Given also that the additional cost of the program is less than $2 per student per year, this suggests that the program is highly cost-effective.

Moreover, the fact that the EFA is a generic program, which was delivered at scale, with minimal support provided to schools, suggests that it has the potential to significantly increase student achievement, at scale, in a sustainable way, within existing resources constraints.

Conclusion

The main argument of this chapter is that the most productive way of implementing formative assessment at scale is to recognize that formative assessment has both generic

and domain-specific elements. The chapters in this volume provide detailed, important, and useful guidance on the practical issues that need to be considered when implementing classroom formative assessment. Formative assessment in English is different from formative assessment in mathematics, which is in turn different from formative assessment in science. While some of these differences are more to do with the way that the subjects have been defined in school curricula rather than anything inherent in the discipline, supporting teachers in their development of classroom formative assessment has to recognize the realities of their classrooms. However, to make students' school experiences more coherent, to allow schools to create a shared language of description in order to improve communication between teachers, it is also important to recognize that many aspects of formative assessment—including its definition—can and should be addressed generically.

Moreover, as the evaluation of the EFA project above shows, there is now clear evidence that generic approaches to teacher professional development can be effective. However, the available empirical evidence does also suggest that approaches to formative assessment that take into account the particular issues involved in implementing formative assessment in different disciplines and domains do produce somewhat larger effect sizes. Obviously, more research will be needed to refine the somewhat limited evidence of effectiveness that has been discussed here, but it does seem to indicate that attempts to harness the power of formative assessment need to recognize that it is both subject specific and generic.

Perhaps the fundamental question, in taking this work forward, is whether it is more fruitful, in the development of formative assessment practice, to work "top-down" or "bottom-up." In other words, should we start with generic approaches to formative assessment, and explore how these approaches can best be implemented in particular domains? Or would it be better to begin with detailed conceptualizations of domains, and then select particular formative assessment strategies that would appear to be especially relevant for that domain? Further work in articulating the differences between these two approaches, and their implications for practice, would allow the strengths and weakness of the two approaches to be investigated empirically. What the available evidence does show is that formative assessment represents an extremely powerful focus for effective, scalable, teacher professional development. While finding the optimum balance between generic and domain-specific approaches is likely to be challenging, the potential benefits for learners suggest that it is likely to be a highly productive focus for future work.

References

Assessment Reform Group. (2002). *Assessment for learning: Ten research-based principles to guide classroom practice*. Cambridge, UK: University of Cambridge Faculty of Education.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York, NY: Holt, Rinehart & Winston.

Banks, B. (1991). *The KMP way to learn maths: A history of the early development of the Kent Mathematics Project*. Maidstone, UK: Bertram Banks.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles Policy and Practice, 18*(1), 5-25.

Black, P., & Harrison, C. (2002). *Science inside the black box: Assessment for learning in the science classroom*. London, UK: King's College London Department of Education and Professional Studies.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham, UK: Open University Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8-21.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.

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

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

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means* (Vol. 68(2), pp. 26-50). Chicago, IL: University of Chicago Press.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289–328.

Broadfoot, P. M., Daugherty, R., Gardner, J., Gipps, C. V., Harlen, W., James, M., & Stobart, G. (1999). *Assessment for learning: Beyond the black box*. Cambridge, UK: University of Cambridge School of Education.

Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Cambridge, MA: Belknap Press.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1491–1494.

Carroll, J. B. (1963). A model for school learning. *Teachers College Record, 64*(8), 723–733.

Chen, F., Lui, A. M., Andrade, H., Valle, C., & Mir, H. (2017). Criteria-referenced formative assessment in the arts. *Educational Assessment, Evaluation and Accountability, 29*(3), 297-314. doi: 10.1007/s11092-017-9259-z

Claxton, G. L. (1995). What kind of learning does self-assessment drive? Developing a 'nose' for quality: Comments on Klenowski. *Assessment in Education: principles, policy and practice, 2*(3), 339-343.

Common Core State Standards Initiative. (2010). *Common core state standards for mathematics*. Washington, DC: National Governors Association/Council of Chief State School Officers.

Cooper, H., Nye, B. A., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*(3), 227-268.

Council of Chief State School Officers. (2008). *Attributes of effective formative assessment*. Washington, DC: Council of Chief State School Officers.

Council of Chief State School Officers. (2018). *Revising the definition of formative assessment*. Washington, DC: Council of Chief State School Officers.

Cox, M., & Webb, M. (2007). *Information and communication technology inside the black box: Assessment for learning in the ICT classroom*. London, UK: NFER-Nelson.

Cronbach, L. J. (1963). Course improvements through evaluation. *Teachers College Record, 64*(8), 672-683.

Dann, R. (2002). *Promoting assessment as learning: improving the learning process*. London, UK: RoutledgeFalmer.

Driver, R., & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students. *Studies in Science Education, 5*(1), 61-84. doi: 10.1080/03057267808559857

DuFour, R. (2007, July 30). Common formative assessments. *All Things PLC.* Retrieved November 25, 2018, from http://www.allthingsplc.info/blog/view/14/common-formative-assessments

Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin.

Earl, L. M., & Katz, S. (2006). *Rethinking classroom assessment with purpose in mind: Assessment for learning, assessment as learning, assessment of learning*. Winnipeg, MB: Manitoba Education, Citizenship and Youth.

Edgar, H. C. (1915). *Sentence analysis by diagram: A handbook for the rapid review of English syntax*. New York, NY: Newson & Company.

Education Endowment Foundation. (2018). Teaching and learning toolkit. Retrieved January 8, 2018, from https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit

Flew, A. (1975). *Thinking about thinking: Do I sincerely want to be right?* London, UK: Fontana.

Gobert, J. D., O'Dwyer, L., Horwitz, P., Buckley, B. C., Levy, S. T., & Wilensky, U. (2011). Examining the relationship between students' understanding of the nature of models and conceptual learning in biology, physics, and chemistry. *International Journal of Science Education, 33*(5), 653-684

. doi: 10.1080/09500691003720671

Goe, L., & Bridgeman, B. (2006). *Effects of Focus on Standards on academic performance*. Princeton, NJ: Educational Testing Service.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271.

Harrison, C., & Howard, S. (2009). *Inside the primary black box: Assessment for learning in primary and early years classrooms*. London, UK: NFER-Nelson.

Hart, K. M. (Ed.). (1981). *Children's understanding of mathematics: 11-16*. London, UK: John Murray.

Herman, J., Matrundola, D. L. T., Epstein, S., Leon, S., Dai, Y., Reber, S., & Choi, K. (2015). *The implementation and effects of the mathematics design collaborative (MDC): Early findings from Kentucky ninth-grade Algebra 1 courses*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). *Empirical benchmarks for interpreting effect sizes in research*. New York, NY: MDRC.

Hodgen, J., & Wiliam, D. (2006). *Mathematics inside the black box: Assessment for learning in the mathematics classroom*. London, UK: NFER-Nelson.

Jackson, C. K., & Makarin, A. (2016). *Simplifying teaching: A field experiment with online "off-the-shelf" lessons*. Cambridge, MA: National Bureau of Economic Research.

Jones, J., & Wiliam, D. (2007). *Modern foreign languages inside the black box: Assessment for learning in the modern foreign languages classroom*. London, UK: Granada.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Yonkers-on-Hudson, NY: World Book Company.

Kingston, N. M., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Kingston, N. M., & Nash, B. (2015). Erratum. *Educational Measurement: Issues and Practice, 34*(1), 55.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86.

Lambert, D., & Weeden, P. (2006). *Geography inside the black box*. London, UK: NFER-Nelson.

Lawton, D. L. (1975). *Class, culture and the curriculum*. London, UK: Routledge and Kegan Paul.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute-by-minute and day-by-day. *Educational Leadership, 63*(3), 18-24.

Leahy, S., & Wiliam, D. (2013). *Embedding formative assessment*. London, UK: Specialist Schools and Academies Trust.

Marshall, B. (2000). A rough guide to English teachers. *English in Education, 34*(1), 24-41.

Marshall, B., & Wiliam, D. (2006). *English inside the black box: Assessment for learning in the English classroom*. London, UK: NFER-Nelson.

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Duit, F. M. Goldberg & H. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies (Proceedings of an international workshop held at the University of Bremen, March 4-8, 1991)* (pp. 110-128). Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence centered design* (Vol. RR-03-16). Princeton, NJ: Educational Testing Service.

Moreland, J., Jones, A., & Barlex, D. (2008). *Design and technology inside the black box: Assessment for learning in the design and technology classroom*. London, UK: NFER-Nelson.

National Assessment of Educational Progress. (2013). *Trends in academic progress: Reading 1971–2012, mathematics 1973–2012*. Washington, DC: United States Department of Education.

Nuthall, G. (2007). *The hidden lives of learners*. Wellington, NZ: New Zealand Council for Educational Research.

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 results: Policies and practices for successful schools* (Vol. 2). Paris, France: Organisation for Economic Co-operation and Development.

Parkhurst, H. (1922). *Education on the Dalton Plan*. London, UK: G. Bell and Sons, Ltd.

Popham, W. J. (2006). Phony formative assessments: Buyer beware! *Educational Leadership, 64*(3), 86-87.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1), 4-13.

Reiser, R. A. (1986). Instructional technology: A history. In R. M. Gagné (Ed.), *Instructional technology: Foundations* (pp. 11-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education, 17*(1), 1-24.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. doi: 10.1016/j.tics.2010.09.003

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal, 50*(5), 1020-1049. doi: 10.3102/0002831213477680

Saunders, W. M., Goldenberg, C. N., & Gallimore, R. (2009). Increasing achievement by focusing grade level teams on improving classroom learning: A prospective, quasi-experimental study of title 1 schools. *American Educational Research Journal, 46*(4), 1006-1033.

Scriven, M. (1963). *The methodology of evaluation*. Lafayette, IN: Purdue University.

Sfard, A. (1998). On two metaphors for learning and on the dangers of choosing just one. *Educational Researcher, 27*(2), 4-13.

Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning* (pp. 279-303). Mahwah, NJ: Lawrence Erlbaum Associates.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large‐scale assessment. *Educational Measurement: Issues and Practice, 37*(1), 21-34. doi: doi:10.1111/emip.12189

Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H., & Anders, J. (2018). *Embedding Formative Assessment: Evaluation report and executive summary*. London, UK: Education Endowment Foundation.

Stiggins, R. (2010). Essential formative assessment competencies for teachers and school leaders. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 233-250). New York, NY: Taylor & Francis.

Sweller, J., Kalyuga, S., & Ayres, P. (2011). *Cognitive load theory*. New York, NY: Springer.

Tan, F., Teng, E., Tan, J., & Peng, Y. W. (2014, May). *Holistic assessment implementation in Singapore primary schools part II: Developing teacher assessment capacity to improve student learning*. Paper presented at the Annual meeting of the International Association for Education Assessment,, Singapore.

Thompson, M., & Wiliam, D. (2008). Tight but loose: A conceptual framework for scaling up school reforms. In E. C. Wylie (Ed.), *Tight but loose: Scaling up teacher professional development in diverse contexts* (Vol. RR-08-29, pp. 1-44). Princeton, NJ: Educational Testing Service.

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. Portland, OR: North West Evaluation Association.

Washburne, C. (1941). *A living philosophy of education*. Chicago, IL: University of Chicago Press.

Wiliam, D. (1992). Special needs and the distribution of attainment in the national curriculum. *British Journal of Educational Psychology, 62*, 397-403.

Wiliam, D., & Leahy, S. (2015). *Embedding formative assessment: Practical techniques for K-12 classrooms*. West Palm Beach, FL: Learning Sciences International.

Wiliam, D., Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: impact on student achievement. *Assessment in Education: Principles Policy and Practice, 11*(1), 49-65.

Wiliam, D., & Thompson, M. (2008). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning* (pp. 53-82). Mahwah, NJ: Lawrence Erlbaum Associates.

---

[1] This might seem like a strong claim, but it is consistent with some other estimates of the differences in rates of learning between high achievers and low achievers (see, for example, Wiliam, 1992)

[2] The use of the word "sound" here in effect renders the statement unfalsifiable, since any exceptions can be defined as unsound—a debating technique that that Anthony Flew (1975) describes as a "no true Scotsman" move (p. 47).

[3] The study was pre-registered as ISRCTN ISRCTN10973392 at https://www.isrctn.com/ISRCTN10973392.

[4] This is, in fact, a slight simplification. The Attainment 8 score is based on the student's grade in math, English language, the three best grades in the subjects included in the English baccalaureate (science, foreign languages, history, geography), and their three best GCSE or equivalent grades in other subjects, with the grades for math and English being double-weighted.