DOI:
10.1080/0969594X.2018.1441807

Document Version
Peer reviewed version

Classroom assessment and pedagogy

Paul Black and Dylan Wiliam

1 Introduction

Our 1998 article on formative assessment was a review of 250 research articles which were relevant to practices in formative assessment. That review did not start from a pre-defined theoretical base; it was basically pragmatic and did not aim to formulate any over-arching theory—it might be regarded as what has been termed a 'configurative review' (Sandelowski, Voils, Leeman, & Crandell, 2011). It did however encourage widespread interest in the topic. In subsequent work (Black et al. 2002, 2003) the practical implementations of various formative assessment practices were explored with and by teachers: the findings of this work did produce explicit guidance to help teachers develop their use of formative assessment. However, one conclusion of the 1998 article was that "building coherent theories, adequate descriptions , and firmly grounded guides to practice, for formative assessment is a formidable undertaking" (Black & Wiliam, 1998, p.61).

In responding to that challenge, we attempted a consideration of factors that should form part of any theory of formative assessment (Black & Wiliam, 2009), but did conclude in that publication that "there is ample room for such considerations" (p.28). In this article we address one aspect of this lack of closure: that, as noted by Phillippe Perrenoud in his response to our earlier review of research into formative assessment (Black & Wiliam, 1998), any adequate theory of formative assessment needs to be embedded in a wider theoretical field (Perrenoud, 1998). Specifically, we propose that formative assessment cannot be fully understood except within the context of a theory of pedagogy. As evidence that the business is, indeed, unfinished, we would point to the fact that much, and perhaps most, of the scholarly literature in the area of pedagogy says little or nothing about assessment.

In part this may be because most writers on pedagogy have been concerned to focus on the social, cultural and political context within which schooling in general, and the curriculum in particular, are framed. In part, it may also be the result of a distinction drawn by some between *pedagogy* and *instruction*, in which only the latter is concerned with the day-to-day work of teachers in classrooms.

However, we do not believe that it makes sense to assume that assessment is merely a secondary influence on classroom practice, especially in those countries, such as Australia, England, and the USA, that emphasize holding schools and teachers accountable for the scores their students achieve on standardized tests, and where, as a result, student performance on these tests influences all aspects of the work of schools.

Of course a clear distinction could still be maintained by arguing that formative assessment is about the details of instruction, and that it is only summative assessment that should feature in the broader study of pedagogy. In response, we would argue that such a position is

unhelpful, because discussing formative aspects of assessment as a part of instruction, with summative aspects of assessment considered as part of pedagogy, inevitably makes it more difficult to think about assessment in an integrated way.

This may seem to be a rather minor issue, but, as a result of our work with teachers, we have become convinced that any approach to the improvement of classroom practice that is focused on assessment must deal with *all* aspects of assessment in an integrated way. For example, while it is possible for researchers to make clear theoretical distinctions between formative and summative aspects of assessment, for teachers such distinctions are at best unhelpful, and may even be counter-productive. For us, it is striking that, in every jurisdiction in which we have worked with teachers on the development of formative assessment practice, teachers have said that the practices we are advocating cannot be used because they are under pressure to raise their students' test scores, and specifically, as a result, they have to 'teach to the test'. This is a particularly ironic finding given the evidence of the impact of formative assessment practices on student achievement on standardized tests (Wiliam et al., 2004; Kingston & Nash, 2011; 2015). It is therefore, we believe, hard to see how further development of formative assessment can take place without clarifying the relationships between the formative and the summative aspects of teachers' work, whilst developing more nuanced understanding of the desirable, and indeed the possible, relationships between them. The danger in all this, of course, is that by widening the focus of our work on assessment, the focus becomes too broad, so that a theory of formative assessment becomes a rather weak "theory of everything."

For example, some authors have proposed adding, to the ideas of assessment *of* learning and assessment *for* learning, a third aspect of assessment: assessment *as* learning. Now of course the idea that students should be learning something while they are being assessed is very attractive. Getting stable estimates of student achievement from authentic tasks requires a considerable amount of time to be devoted to assessment (see, e.g., Linn, 1994), which is hard to justify if students are not learning anything while undertaking those assessment tasks. An assessment activity in which students learn something, is, all other things being equal, preferable to one in which they do not. However, if the term formative assessment is to be helpful, we believe strongly that it is necessary to keep a clear focus on the idea of formative assessment as *assessment*—a point also made forcefully by Bennett (2011).

Particularly in Canada and Australia, the term "assessment as learning" (Hayward, 2015) has been used in a more specific sense; to describe the role of students in monitoring and directing their own learning (e.g., Earl, 2003; NSW BSTES, 2012). The term 'assessment for learning' is then used to describe the process by which teachers use assessment evidence to inform their teaching, and 'assessment of learning' refers to the use of assessment to determine the extent to which students have achieved intended learning outcomes. However, this approach does not provide any straightforward way of integrating the role that peers have to play in supporting learning, and so we believe that these various "prepositional permutations" (Popham, 2005) serve to obscure rather than illuminate the important processes.

Lee Cronbach (1971) proposed that an assessment is, at its heart, a procedure for making inferences: "One validates, not a test, but an *interpretation of data arising from a specified*

procedure" (p. 447, emphasis in original). An educational assessment is therefore a procedure for making inferences about student learning. Learners engage in tasks, which generate data. These *data* become *evidence* when they are used to in support of particular claims: "A datum becomes evidence in some analytic problem when its relevance to one or more hypotheses being considered is established...evidence is relevant on some hypothesis if it either increases or decreases the likeliness of the hypothesis" (Schum, 1987 p. 16). In other words, "Educational assessment is at heart an exercise in evidentiary reasoning." (Mislevy & Riconscente, 2005 p. iv).

From such a perspective, the distinction between formative and summative becomes *a distinction in the kinds of inferences being drawn from assessment outcomes*. Where the inferences relate to the status of the student, or about their future potential, then the assessment is functioning summatively. Where the inferences relate to the kinds of actions that would best help the student learn, then the assessment is functioning formatively. As we have argued elsewhere (see, e.g., Wiliam & Black, 1996), one immediate consequence of this perspective is that assessments themselves cannot be formative or summative. If we look at a student's responses to a test of multiplication facts, we might conclude that the student knows approximately 80% of his multiplication facts (a summative inference) or we might conclude that the student appears to have particular difficulties with his seven times tables (a formative inference). The same assessment instrument, and indeed, the same assessment outcomes, can be used both summatively and formatively, although such uses may not be equally valid: "Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test 'is valid'" (Cronbach, 1971 p. 447); this point is further emphasised in the review by Newton who stated that "It therefore makes sense to speak of the validity of the assessment-based decision-making procedure, which is underwritten by an argument that the assessment procedure can be used to measure the attribute entailed by that decision" (Newton, 2012, p.25).

Cutting across the distinction between summative and formative functions of assessment are issues about who decides on the assessment, where the assessment takes place, and how the students' work is scored. In this paper, we use the term classroom assessment to describe those assessments where the main decisions about what gets assessed, how the students will be assessed, and the scoring of the students' responses, is undertaken by those who are responsible for teaching the same students. Externally produced assessments could be regarded as classroom assessments if the decision about whether to use those assessments, how to administer them, and how to score them rested with the teacher.

The focus of this paper, then, is to embed both formative and summative aspects of classroom assessment within a wider theoretical framework, as recommended by Perrenoud, while maintaining a clear focus on classroom assessment *as assessment*—a procedure for making inferences. The purpose of the assessment may be broken up into a collection of many different aspects—some of these inferences will be about the stage of learning that a student has reached, and other inferences will be related to the courses of action that would best advance a student's learning—but it is not obvious that this collection either could be, or should be, neatly divided into two distinct groups.

This analysis presented here will, we hope, serve two purposes. The first is to support further development of studies of pedagogy that might be better than existing studies in giving due emphasis to the importance of assessment, linking it into the broader and complex overview which such studies should seek to achieve. The second is to apply that framework to exploring the relationship between the formative and the summative purposes, to suggest that this relationship could be—and should be—one of helpful overlap—and perhaps even mutual support—and that it is natural that it should be so.

However, these are limited claims, in that we are not claiming that its work could be the sole basis for a revised theory of pedagogy. For example, this work could be seen as one contribution to the "ongoing attempt to understand the outer limits and inner constraints of forms of pedagogic communication, their practices of transmission and acquisition, and the conditions of their change" (Bernstein, 1990, p.9). An attempt to identify the key points of this contribution is included in section 7 (Conclusion) below.

2 Pedagogy and/or Instruction.

The terms *pedagogy* and *instruction* are used widely in discussions of educational processes, although the extent and nature of their use varies from country to country. Writing in 1981, Brian Simon pointed out that neither the term *pedagogy*, nor the ideas it conveyed, were commonly part of educational discussions in England, with teachers planning and explaining their teaching in largely pragmatic or ideological terms (Simon, 1981). Almost two decades later, Watkins and Mortimore (1999) echoed this observation, but also pointed out that both the term and the associated ideas were beginning to gain greater currency, with different approaches reflecting different emphases amongst the many aspects of teaching and learning. Many of these approaches, such as that of Paulo Freire (1970), qualified the term pedagogy with adjectives such as critical, conflict, liberatory and gender, all of which highlight the political function of pedagogy.

In contrast, others have suggested that the term pedagogy should be used narrowly, to describe simply "competence, excellence and failure in teaching methods" (Anthea Millett, quoted in Alexander, 2008 p. 45), treating teachers, in Lawrence Stenhouse's memorable phrase, "as a kind of 'intellectual navvy', working on a site plan simplified so that people know exactly where to dig their trenches without knowing why" (Stenhouse, 1985 p. 85).

Whatever the relative merits of broad and narrow definitions of pedagogy, a narrow definition of pedagogy would exclude assessment entirely, or at the very least, even if it did include some aspects of assessment, it would exclude others, and thus make it impossible to deal with assessment in an integrated way. For the purposes of this paper, we therefore define pedagogy broadly, and specifically as "the act and discourse of teaching" (Alexander, 2004 p. 8) thus explicitly including curriculum and assessment, Specifically, we adopt Alexander's (2008) definition:

> *pedagogy is the act of teaching together with its attendant discourse of educational theories, values, evidence and justifications. It is what one needs to know, and the skills one needs to command, in order to make and justify the many different kinds of decision of which teaching is constituted*. Curriculum is just one of its domains, albeit a central one. (p. 47 original emphasis)

Where such a broad definition is adopted, it follows that the term *instruction* must represent some realm that is one component of the broader realm of pedagogy. Shulman (1999) adopted a similar schema, as did Hallam and Ireson (1999) in describing pedagogy as "those factors affecting the processes of teaching and learning and the inter-relationships between them" (p.78). Very different definitions, which we shall not adopt here, are chosen, for example, by Bruner. In *Toward a theory of instruction* (1966) , stating that " A theory of instruction is *prescriptive* in that it sets forth rules concerning the most effective way in achieving knowledge or skill", contrasting it with theories of learning which are descriptive (p.40). This variety in the definitions does not seem to us to reflect any fundamental difference between those who use these terms in these different ways.

In terms of the focus of this paper, what is notable in almost all of this literature on pedagogy is that assessment receives scant attention (Black & Atkin, 2014), and it is this lack of attention to assessment that we seek to address. Our argument is that any examination of pedagogy that does not take into account the various kinds of assessment processes that are used in educational settings—or does not explicitly analyse such processes as assessment—can at best provide only a partial explanation of what is going on. Our particular focus is on the context of classroom teaching and learning, and, in the remainder of this paper we consider how to model pedagogy and instruction in a way that includes assessment, and then show how the literature contributes to the various components that are relevant here, even although the studies quoted have paid very little attention to assessment as such.

3 Models for pedagogy and instruction

In his widely read *Principles of curriculum and instruction* (still in print more than six decades after its first publication), Ralph Tyler proposed that the curriculum should be seen as a means to an end, rather than an end in itself. He proposed an instructional model focused around "four fundamental questions which must be answered in developing any curriculum and plan of instruction" (Tyler, 1949 p. 1):

1. What educational purposes should the school seek to attain?
2. What educational experiences can be provided that are likely to attain these purposes?
3. How can these educational experiences be effectively organized?
4. How can we determine whether these purposes are being attained?

A similar model, but with slightly different emphases, was proposed by Black (2016):

Step 1 – *Clear aims*
Step 2 – *Planning activities*
Step 3 – *Interaction in dialogue*
Step 4 – *Review of the learning*
Step 5 – *Formal summative assessment*

Whilst schemes of this type, and particularly its first three components, are discussed by many authors, to our knowledge only two (Hallam & Ireson, 1999; Wiske, 1999) set out and discuss all five in a similar sequence. However, whilst these five can be seen to represent

successive stages in the planning and then implementation of any piece of teaching, representing the complex interaction of factors that bear on pedagogy is not straightforward. In addition, whilst its simplicity lies in its representation of a time sequence of decisions, it does not follow that links between these steps are implemented in only one direction; it is likely that there will be cyclic interactions between the components, as, for example, where a step 4 review leads to re-shaping of step 2 leading in turn to a different emphasis in the implementation of step 3.

A more complex model is presented in figure 1. The factors which combine in the formulation of aims are represented in boxes 1, 2, and 3, the planning and design of activities is box 4, implementation and review of the learning are in box 5, with external summative assessment in box 6.

In the next section, we discuss each of these components in turn, in the light of the general literature on pedagogy and of our own previous work in formative assessment. In a subsequent section, we shall take up the issue of assessment for summative purposes.

«Figure 1 about here»

4 Components of the model

The broad determinants of pedagogy and instruction

The literature presents a variety of priorities in respect of the aims of pedagogy. As noted above, many of the stated aims entail commitments about *what* students should learn, as well as how they should learn it, such as the following: "What ultimately counts is the extent to which instruction requires students to think, not just to report someone else's thinking" (Nystrand Gamoran, Kachur, & Prendergast, 1997, p.72).

Other approaches provide guidance about the sequencing of learning activities based on the kinds of cognitive processes that might be expected from students as they get older. Bruner (1966) proposed that activities should be designed to promote progression in thinking—from enactive to iconic to symbolic, while Tyler (1949) emphasized a progression from inductive thinking through deductive thinking to assembling complex arguments from a variety of sources. In her review of programmes aimed at developing thinking skills, McGuinness (2005) identifies the development of metacognition as a core feature common to all of them. Both Bruner and Alexander express related views in different ways, giving priority to theories of learning (Figure 1, box 2), emphasising that the increasingly rapid pace of change means that the ability to continue to learn beyond school will be paramount and arguing that dialogue is essential to developing learners' power of critical review through which they can come to internalise language as an instrument of thought. Clearly, the opportunities for students to internalise language in this way will be greater where students have more opportunities to talk, and so group work and collaborative learning will be essential components of effective learning environments. However, it seems, particularly at secondary school level, the more common everyday priority of teachers is to express aims in terms of the learning of the content of particular subjects (box 3). This is partly a 'content' argument, but it is also important to note that an important role for education is to develop the

'disciplinary habits of mind' peculiar to different subject disciplines. A specific example is the analysis by Leach and Scott (2005) about the design of science teaching; they argue that the various learning demands are as much epistemological as psychological, whilst difficulties with ontology may also be important. However, all would, we believe, agree that the content topics chosen must be central to the subjects involved, and that a key aim should be to develop students' abilities to apply what they have learned in one context to another context (Nuthall, 2007).

The importance of dialogue also contributes to a different aim; to develop, through its emphasis on sociocultural aspects of learning, the habits of collaboration and of working in and through a community. Such arguments link with wider concerns about the development of learners' expectations, and of the transmission through schooling, of learners' awareness of more fundamental concepts of values and of virtue. Tyler points out that the teacher's task must be to relate aims to practice, which calls for clear values and a philosophy of education: given this, his assertion that teachers rarely spell out their objectives are, if still true, a cause for concern.

Whilst these different aims may compete for priority, they are not mutually exclusive. Some may have direct effects on the activities planned (box 4), others in the way activities are implemented (box 5), and so on.

Planning and design

All the authors quoted above specify several elements of planning, and many of these are common to a number of authors. However, the task in practice is to organise these elements and then, as Tyler puts it, to weave the threads together.

Planning has to take places at several levels from the overall and general to the detailed and particular. At a high level of generality is the need for an organising scheme, structured so that the aims, in terms of learning aims and content aims, may serve to guide specific activities. A structure should guide a sequence of activities, so that each successive topic stimulates recall of previous work, connecting to it and thereby expanding the range of the learning. Particular attention to developing the learning power of the students would be a component in such a sequence. At a more detailed level, as Hilda Taba (1962) pointed out, each specific activity ought to be so designed that it engages attention, motivates and presents a challenge to which the learners can respond with high probability of success, so securing intrinsic reward. Good activities can engage and develop interest, which is both of value for the particular lesson and for the larger aim of building up the learner's commitment to learning.

Two features appear to us to be especially important in closing the gap between these general principles and the formulation of a specific lesson activity. One is that a task must be so fashioned and presented that, at the outset, it engages the learners in a way that it will help elicit evidence of their understanding. In other words, rather than the teacher presenting a summary of previous learning and assuming that this level of understanding is shared, the teacher elicits evidence from the class, and, on the basis of that evidence, makes inferences about what next steps in instruction are likely to be most effective.

The second is that, given such elicitation, the challenge of the task can be developed to involve learners actively in developing that understanding *in the direction planned*, notably through their engagement in dialogue with the teacher and with one another. In this context, it is worth noting that one of the important differences between experts and novice teachers is the way that they respond to unexpected student responses or questions. Novice teachers tend to adopt one of two opposite courses of action. Sometimes novices pursue the unexpected responses in detail, losing track of the original intention of the lesson, while at other times they ignore the students' responses entirely, sticking to their original plan for the lesson. Experts, on the other hand, rapidly appraise the content of the unexpected response or question and decide whether it is likely to be helpful in advancing the intended learning of the whole group (Berliner, 1994).

These two priorities are very similar to the two main priorities stated by Alexander. Reflection on experience and foresight are needed if such formulation is to lead to effective action, and, given the complexity and context sensitivity of what is required, the practical knowledge and skills of the expert teacher cannot be captured by a check-list of rules. Thus, it is not surprising that scholarly books and articles do not present specific examples. Such simple recipes as the three-part lesson (DfE 2010) may inhibit rather than enrich (NCETM, 2007), but numerous and more varied examples of lesson plans are available on various web sites (see, for example, TeacherNet 2010).

In such cases, the importance of assessment as a key component in the regulation of learning processes (Allal, 1988)—of "keeping learning on track" (Wiliam, 2007)—is clear. Instruction is designed in such a way that teachers gain evidence about students' capabilities, and also so that this evidence is used to adjust instruction to better meet the students' learning needs—what might be termed a pedagogy of responsiveness (Shulman, 2005; Wiliam, 2016). In particular, a focus on the quality of the evidence that teachers have to support their instructional decisions highlights the need for the teacher to get evidence from all students in the class, rather than only those confident enough to volunteer their views—a pedagogy of engagement (Shulman, 2005; Wiliam 2016).

Implementation – teachers' formative assessment

The implementation of teachers' intentions and plans in the classroom is a pivotal feature. Some treatments (e.g. Tyler, 1949; Shulman, 1999) say little about this stage. However, studies of discourse, of formative assessment, and of dialogue, have provided rich resources for detailed exploration of this area. Moore (2000) linked the findings of the discourse literature to a more general discussion of pedagogy; Black, Harrison, Lee, Marshall, & Wiliam (2003) explored the application of findings from research studies of formative assessment to practices in a range of classrooms, whilst Alexander (2008) has drawn on his landmark study of classrooms in several countries (Alexander, 2001) and subsequent research in UK classrooms, to summarise findings about classroom dialogue.

Nystrand's phrase "when recitation starts, remembering and guessing supplant thinking" (Nystrand et al., 1997, p.6) identifies a common failing. The range of classroom practices is described, by Alexander, as a spectrum thus – rote, recitation, instruction by exposition,

discussion and dialogue. There is ample evidence that much practice, at least in the USA (Applebee, Langer, Nystrand, & Gamoran, 2003) and the UK (Smith, Hardman, Wall, & Mroz, 2004) pays little attention to the discussion and dialogue end of this spectrum. Yet interactive dialogue, with its essentially contingent nature, which is a strong component of formative assessment practices, does lead to more effective learning.

It is clear that plans that include the intention of engaging pupils in purposeful learning will succeed only if the teacher can open up and then steer a learning dialogue. This is a delicate task, involving elicitation of pupils' contributions, responding in a way that seeks to use, and so to value, even the most bizarre of pupil suggestions, and conducting the discussion in such a way that the progress is, and is seen to be, partly the responsibility of the pupils, and yet is kept on track with the larger aims in view. Such steering should lead to interpretations of the outcomes that would seek to transform them in the light of a broader perspective, linking them to earlier learning and opening up the next steps. Key indicators of the quality of such dialogue are the use by pupils of such terms as 'think', 'because', 'would' and should, and whether their contributions are in the form of single words and short phrases, or sentences, of even paragraphs. A helpful typology of teachers' questions in classroom dialogue can be found in Kawalkar and Vijapurkar (2013).

There are many pathologies: a common case is when a teacher, faced with a bizarre response, can say no more than 'not what I was thinking of'; another is the tendency for a dialogue to regress to recitation in a well-meaning attempt to 'put things right' or for a teacher to 'recast' a student's response in a way that makes it correct even though there is substantial evidence that such approaches are ineffective (Lyster & Ranta, 1997); a third is the use of what Alexander describes as 'phatic praise'. Indeed praise that does not take seriously and build on the potential of a response, to explore and challenge a learner's thinking, misses learning opportunities and might even do positive harm. Of course, defects in the design and presentation of the activity will make all of this difficult: a colleague reported a class in which, after 15 minutes of tangled discussion, the teacher had the courage to say to the class "This isn't working is it?": the students agreed, and the teacher started again in a different way. All of these issues are explored in more detail in the analysis of interactive formative assessment in our earlier paper (Black & Wiliam, 2009).

The analysis of Wiliam and Thompson, (2007), shows how the various activities through which formative assessment is implemented can be conceptualized as consisting of a map of five key strategies (figure 2):

<<Figure 2 about here>>

The metaphor of engineering, in strategy 2, reflects the complex task of the teacher in assembling and co-ordinating a range of possible tactics to serve these strategies, whilst the fourth and fifth in this list draw attention to the need to articulate whole class teaching with other work, notably group discussion, peer-assessment, and the formative use of such methods in the marking of written work, including outcomes of tests. Thus, whilst it is easy to see oral classroom dialogue as the core of formative assessment, formative interaction is also realised in relation to various types of written work, albeit in asynchronous form. Thus,

written or other productions produced in the last stage of an activity, whether as classwork or as homework, set up opportunities for further formative feedback.

The framing of strategy 3 as 'feedback that moves learning forward' draws explicit attention to the idea that feedback should, for the most part, be focused on improving students' performance on tasks they have not yet attempted, rather than on rendering judgements on the adequacy of previous performance. Moreover, this formulation highlights the need to focus on the effect of feedback rather than its intent. As Dweck (2000) has shown feedback in the form of grades can serve to focus attention on comparison with others, and can also lead to a belief that ability is fixed, rather than malleable. Feedback in the form of comments about how to improve is more likely to send the message that ability can be improved, so that feedback is welcomed as directing that improvement.

The well-established findings on the effects of feedback highlight the importance of one-to-one exchanges of the teacher with each pupil, either in written (but comment-only) marking, or orally. An alternative use is to set up peer-assessment groups in which pupils compare and rank one another's work, thereby developing further their understanding of the criteria for quality in the expression of their thinking about the issues involved. However, as is clear from the research of Robert Slavin (e.g., Slavin, Hurley & Chamberlain, 2003) and the Johnson brothers (Johnson & Johnson, 2009) such peer interaction is likely to be successful only when there are group goals (so that students are working as a group, not just in a group) and that students are individually accountable to the group for the quality of their contributions. However, research studies of the work of pupil-groups in classrooms have shown that the peer-interactions are not effective unless pupils are given some training to guide them to emphasise reasons rather than assertions and to offset the tendency of some to dominate the discussion (Baines et al., 2009, Mercer et al., 2004).

In terms of the goal of this paper, the key point here is that the contingent nature of the learning process—whether in dialogue, when a teacher writes comments on a student's work, or when a student gets feedback from a peer—means that the elicitation, identification and interpretation of evidence is an indispensable component of effective instruction. Assessment *as assessment* is at the heart of effective instruction.

Implementation – teachers' summative assessment

Many studies of pedagogy and instruction play little attention to assessment and testing of pupils. One reason for this is that earlier authors were concerned with assessment seen as evaluation. Tyler discusses the several levels of evaluation, from the summative review of a teaching programme to the evaluation of the curriculum as a whole, whilst both Bruner and Shulman consider the value of evaluative feedback for teachers in helping them reflect on their work. However, there is little or no consideration of evaluative feedback to the pupils. For example, the terms test, formative, summative, examinations, do not appear in the index of the book by Bruner (1966) and in the index for the collection edited by Mortimore (1999), whilst the term assessment, whilst appearing in these and in others (e.g., Moore, 2000), appears as one of several headings in lists, rather than as embedded in substantial discussions of assessment.

As yet another example, Alexander (2008) lists the core acts of teaching as task, activity, interaction and assessment (p.49), but in the subsequent discussion, the last of these is given almost no attention, except in a discussion of the harmful effects of teaching to the test which follow from narrowly designed accountability systems. He argues that in such teaching, the learning power of the pupils is undermined – they are victims rather than agents. He traces this fault to the deterministic and fatalistic ideas about learning which are rooted in Anglo-Saxon cultures, contrasting this with beliefs, common in continental Europe, in human perfectability and in the power of external agents to improve the learning power of pupils.

One exception to this neglect of assessment is in the discussion by Wiske (1999), drawing on findings from the Teaching for Understanding Project in the USA. She starts by drawing attention to the training of athletes, for whom the performance as such is open to direct observation. **Their education is designed to help them to understand the features that limit performance, so that they can then compare their own performance with that of successful peers and thereby understand the criteria of excellence. Learners can thus become able to judge their own achievements and limitations, so that summative assessment by others may be superfluous.** Wiliam and Leahy (2015) point out that a similar focus is often seen in the practice of instrumental music teachers (pp. 178-179), and Wiske suggests that a similar approach could be applied in academic subjects, but that to do so would require that learners have a clear understanding of the performances which they are expected to be able to accomplish—a point forcefully made by Royce Sadler (1989):

> The indispensable conditions for improvement are that the student comes to hold a concept of quality roughly similar to that held by the teacher, is able to monitor continuously the quality of what is being produced during the act of production itself, and has a repertoire of alternative moves or strategies from which to draw at any given point. In other words, students have to be able to judge the quality of what they are producing and be able to regulate what they are doing during the doing of it. (p. 121)

Wiske then takes the argument further to envisage that a strategy should be based on use of portfolios, collected from frequent assessments, drawing on multiple sources (teacher, peers, self), linked to progress, and so designed that the collection covers all of the subject criteria. However, she recognise the difficulty of making the policy changes so that such a process could lead smoothly to summative assessment: "These shifts run counter to the norms in many classrooms and require both students and teachers to take on new roles and relationships" (p.242). It is also worth noting that Wiske, like many others who advocate the use of portfolios, does not address the technical difficulties in producing reliable evidence of student achievement across a complex domain with portfolios (see, for example, Koretz, 1998).

A second exception is in our earlier work on the formative use of summative tests (Black et al. 2003). Here, summative tests were used by teachers to generate feedback in the same way as other written work. In other words, they were treated as similar to pieces of written homework for which peer marking was designed to help pupils' self-reflection, and feedback was designed to improve performance rather than to judge or grade. The tests themselves were clearly designed to serve a summative purpose, but in this case were used to support learning, in two ways. The first is that tests designed for summative purposes have a role to

play in clarifying what domain descriptions actually mean. While different teachers might disagree about what it means to be able to "apply the gas laws", an assessment of this skill removes the ambiguity: assessments operationalize constructs (Wiliam, 2010). The second is that assessments can be used as learning activities, as when a teacher asks students to complete a test individually, under test conditions, but then students work in teams to produce the best composite response to the test (Wiliam, 2011).

As noted above, the formative-summative distinction rests in the kinds of inferences that are drawn on the basis of assessment outcomes, and although an assessment can be designed to serve both purposes, when priority is be given to one particular purpose, this is likely to influence the design of the assessment.

For this reason it is impossible, and—at least from the teacher's standpoint—probably unhelpful, to draw a clear demarcation between, for example, a piece of homework for which the teachers may both assign and record a mark, and which may also be used to give formative feedback to promote improvement, on the one hand from a class or end-of-topic test for which the work produced is used in these same two ways on the other. What distinguishes teachers' use of assessments for summative purposes is that the aim is to produce a product, whether in the form of a number, or a grade, justified by results of formal tests, or of coursework, or of a combination of the two. Such products, produced at the end of a phase of a learning programme, can serve a range of purposes. This range includes feedback to the teachers and to their schools about the effectiveness of the teaching, recording achievement of each individual pupil in the group of learning tasks involved, and advice to teachers, to pupils and to their parents about decisions on future courses of study and/or career choices. For all of these it is arguable that, in varying degrees, these functions, whilst obviously summative, may also be formative. They may be formative for teachers, in informing changes in their planning and implementation of teaching the same, or other students in the future, and they may also be formative for pupils in informing reflection on the strengths and weaknesses of their achievements in ways that might help them re-direct their energies in future work. Thus, whilst the feedback use of results will not be ruled out, the level at which they may be used to improve learning is different, and these functions of feedback are secondary. It is these differences that can lead to tensions between formative and summative practices.

From the foregoing, it should be clear that we do not believe that there should be a sharp discontinuity between the formative and the summative. In fact there can be and ought to be synergy between them, and such synergy should be seen as the natural healthy state.

Yet in any system where teachers experience the accountability pressures of externally imposed tests, achieving such synergy may well be seen as a mountain to climb: so, for example, the assessment diagnosis for England is that it has a disease arising from a flawed concept of accountability. In a review of assessment systems in many countries we pointed out that:

> Where summative becomes the hand-maiden of accountability matters can get even worse, for accountability is always in danger of ignoring its effects on learning and thereby of undermining the very aim of improving schooling that it claims to serve. Accountability can only avoid shooting itself in the foot if, in the priorities of

assessment design, it comes *after* learning (Black & Wiliam, 2007, p.10).

This counter-productive effect of accountability arises when, as for example in England, test results are high-stakes for teachers and their schools, whilst the system ignores the effects on learning. Thus, the discussion now centres on boxes 5 and 6 of figure 1 and of the relationship between them.

Ways of overcoming this problem of tension between formative and summative require attention to three components. The one is that teachers and schools should be accountable only for those determinants of their work over which they have control, which requires that some form of (contextual) value-added system be used. This issue will not be discussed here (see Foley & Goldstein, 2011 for a review). A second is that assessment instruments used to serve summative purposes should be so designed that they are supportive of learning, or, at the very least, should not be designed in such a way that is likely to significantly harm learning. The third is that teachers' should take responsibility for serving the summative purposes, or at least play an active part in meeting this responsibility. In the next section, we shall argue that these second and third components are inter-dependent.

5 Summative assessment by teachers

Is the contribution of teachers' findings to summative assessments important?

Over the past few decades, there has been an increasing concern that assessment systems used in schools should support a broader range of inferences about student capabilities than has been the case hitherto, partly because of changes in the kinds of skills needed in employment, but also to reflect the broader aims of education. For example, recent debates in the EU Directorate have identified the importance of developing 'key competences', and are exploring the problems of reflecting these in national assessment systems:

> Key competences are a complex construct to assess: they combine knowledge, skill and attitudes and are underpinned by creativity, problem solving, risk assessment and decision-taking. These dimensions are difficult to capture and yet it is crucial that they are all learned equally. Moreover, in order to respond effectively to the challenges of the modern world, people almost need to deploy key competences in combination. (European Commission Directorate-General for Education and Culture, 2010 p. 35)

Changes in the kinds of inferences that assessments are expected to support have, in turn, highlighted problems with traditional tests and examinations and, specifically, the fact that traditional assessments often do not assess things about which users of assessment information wish to draw conclusions. For example, language competence includes speaking and listening as well as reading and writing, but only the latter are generally assessed (in the psychological jargon, the assessment *under-represents the construct* of language competence). Defenders of traditional assessment point out that measures of performance in reading and writing serve as good proxies for performance in speaking and listening, but that correlation is only likely to hold as long as teachers continue to develop all aspects of language competence. Where the stakes to raise scores are high, teachers and students have

incentives to focus on reading and writing, at the expense of speaking and listening, so that, over time, performance in reading and writing will no longer be a good guide to performance in speaking and listening.

This acknowledgement that assessment use can have social consequences is an explicit feature of the model of validity developed by Samuel Messick. In his landmark chapter on validity in the third edition of *Educational Measurement*, Messick defined validity as follows:

> Validity is an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (Messick, 1989 p. 13)

As a result, some have argued that *all* the social consequences of assessment use should be regarded as aspects of validity. In other words, if tests that are reasonably valid as measures of student achievement are used to hold teachers accountable, then the validity of the tests are somehow called into question. However, Messick explicitly rejected such an argument.

> As has been stressed several times already, it is not that adverse social consequences of test use render the use invalid, but, rather, that adverse social consequences should not be attributable to any source of test invalidity such as construct-irrelevant variance. If the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced—or if the validation process can discount sources of test invalidity as the likely determinants, or at least render them less plausible—then the validity of the test use is not overturned. Adverse social consequences associated with valid test interpretation and use may implicate the attributes validly assessed, to be sure, as they function under the existing social conditions of the applied setting, but they are not in themselves indicative of invalidity. (Messick, 1989 pp. 88-89)

This careful delineation of the extent to which social consequences can be regarded as part of the validity of an assessment is illustrated by current debates over the assessment of practical work in science. In December 2014, the British government began a consultation process over the role of practical assessments in the General Certificate of Secondary Education (GCSE)—the national school leaving examination for England. In March 2015, the government decided that practical work in science would not be assessed directly, but would be assessed through written test items in the terminal examination.

Proponents of this position have argued that organizing practical assessment of laboratory work in science is difficult and expensive. Moreover, the results on such assessments tend to correlate well with other scores on written tests of science achievement. As a result, the argument goes, the scores on written tests would support valid inferences about capabilities in practical science. As J. P. Guilford once said, "In a very general sense, a test is valid for anything with which it correlates" (Guilford, 1946 pp. 428-429). However, when science achievement is measured only through written assessments, particularly in a 'high-stakes' testing environment, then teachers may focus on increasing their students' achievements in the most efficient way possible, and this may well involve doing less practical science. What

is important here is that, at least for Messick, the validity of the science assessment *would* then be jeopardised. The adverse social consequence—students doing less practical work in science —is directly traceable back to the fact that the assessment under-represents the construct of science, at least as defined in the English national curriculum. The important issue here is not that the adverse consequences are bad; it is that they are directly traceable to deficiencies in the assessment.

In the future, it is likely that automated scoring of student responses to essay prompts and constructed-response questions will allow a greater range of student responses to be scored inexpensively, but, for the time being, increasing the validity of student assessment would appear to require the use of information collected by teachers in the course of their teaching. Stanley, MacCann, Gardner, Reynolds and Wild (2009) commented that:

> the teacher is increasingly being seen as the primary assessor in the most important aspects of assessment. The broadening of assessment is based on a view that there are aspects of learning that are important but cannot be adequately assessed by formal external tests. These aspects require human judgment to integrate the many elements of performance behaviours that are required in dealing with authentic assessment tasks. (p. 31)

This specifies a further aspect of learning to add to those components of Box 2 that have been high-lighted above, namely learning by interactive dialogue in both oral and written contexts, by collaboration in group work, and by that promotion of reflection and self-awareness that can be achieved through peer-assessment and self-assessment.

The general point here is that, given the inevitable pressures of high-stakes assessments on teachers' work, priority cannot be given to good learning practices if many of these do not produce those outcomes that will improve the performance of students on these high-stakes assessments. This is why any large scale assessment system which relies solely on formal written tests, set and scored by agencies external to the school, is incompatible with the standards chosen by most education systems as the desired learning outcomes for their students.

Do teachers' potential contribution deserve a role in summative assessment?

A common objection to using teachers' assessments of their own students, especially in high-stakes accountability systems, is that it is akin to 'letting the fox guard the hen-house' because teachers will be tempted to exaggerate their pupils' achievements if it is in their interests to do so. Even where teachers can be trusted to resist such temptation, it is often claimed that it is impossible to secure consistency of teachers' judgements across a local education authority, let alone an entire jurisdiction. Finally, unlike those in other professions, such as medicine or law, teachers do not regulate their own professional standards. Why this might be the case is of course a matter of some debate, but Gardner (2007) suggests that there is "a general milieu of distrust of teacher assessments of students' work" (p. 18). To what extent that distrust is justified is again a matter of debate, but there seems to be little doubt that the typical level of what might be called "assessment literacy" (Stiggins, 1991) in most countries appears to be rather low.

These pessimistic views have been challenged by others, quoting evidence from research studies to spell out the conditions required for such assessments to achieve their potential quality (Harlen 2004; ARG 2006; Daugherty 2010). A more recent survey by Johnson (2011), for England's Office of Qualifications and Examinations Regulation (Ofqual), of the literature on the reliability of teachers' assessments is far more cautious, emphasising both the complexity of the task given that the aim is to achieve validity in a comprehensive way:

> But even for such factual knowledge the assessment challenge increases as we attempt to move away from isolated atomistic assessment exercises in attempts to measure knowledge, abilities and skills more globally. Knowing the date of the Battle of Waterloo is one thing, but this tells us little about the student's broader state of knowledge of that battle, or of that period, or of history in general. Knowing the chemical symbol for mercury gives us no information about how many other chemical symbols the student knows and can correctly attach to the relevant chemicals. It also tells us nothing about the student's knowledge and understanding of chemistry more generally, theoretical or practical (Johnson, 2011 p. 1)

The survey also concludes that the evidence about almost all the procedures involved in the creation of a robust teacher assessment system is inadequate. To take one example—the moderation of teachers' judgments—the conclusion is stark:

> There is no systematic form of moderator training in the GCE/GCSE world, as there is for written test markers, and little is known about inter-moderator consistency at the present time. It would be costly and logistically complex to organise inter-moderator reliability studies on a regular basis for every subject that has a coursework element, but some further research does surely need to be carried out. (op cit, p. 43)

However, despite all this intimidating complexity, it is important to note that the contribution of teachers to formal tests and examinations is only part of the picture. We do need to be concerned with the context of public, local or national assessments, where procedures and instruments have to command public confidence. But there is a broader issue because teachers and their schools have to produce summative assessments of their students on regular, and in many cases on frequent, occasions, particularly when decisions have to be taken at the end of each stage (e.g. each school year) about courses and classes for study in the next stage. To ensure that these decisions are carefully made, students, parents and other teachers need dependable evidence on the progress made by students. Thus validity, reliability and comparability of these measures of progress are needed throughout the work of schools.

Exploring the prospects for positive development

The King's-Oxfordshire Summative Assessment Project (KOSAP; Black, Harrison, Hodgen, Marshall, & Serret, 2010; 2011) illustrates the issues involved in more detail. A *first* finding, which emerged from an initial audit of the internal assessment practices in the three schools involved, was that these did not meet the criteria set out above. Three more conclusions emerged as the research team worked to enhance the

quality of the teachers' summative work. The *second* conclusion was that, assuming the necessity for a system that could produce evidence of work done by students in several tasks (i.e., a portfolio), there had to be some degree of uniformity if moderation of the products, both within and between schools, were to be feasible. This means that it would be necessary to attend to five main features of in each of which threats to both reliability and other aspects of validity can emerge. These are:

(a) Validity of each component of a portfolio: each task or test has to be justified in relation to the aims that it was said to assess.
(b) Agreed conditions for the presentation and guidance under which students would work in producing the various components of a portfolio.
(c) Guidelines about the ways in which each domain is sampled within a portfolio, both by the separate components and by the collection as a whole.
(d) Clear specification of the criteria to which all assessors have to work.
(e) A requirement for comparability of results within and between schools, which would require moderation procedures both within and between schools.

A *third* conclusion, which adds to the findings of Gardner (2007), was that developing the skills and procedures needed would require both training and consistent support if teachers and schools were to develop that confidence and the competence in their assessments. In this regard, the work of Stanley et al. (2009) is useful, in that it describes how the state systems in Queensland and New South Wales provides such training and support.

The *fourth* finding related to the impact on teachers' workload. From the first three conclusions one might infer that such development might be dismissed as unacceptably burdensome. However, this was not borne out. One positive feature was that the moderation processes were seen to have valuable effects on many aspects to the teachers' work:

> …that the moderation and standardisation process was incredibly valuable in ensuring rigour, consistency and confidence with our approach to assessment; that teachers in school were highly motivated by being involved in the process that would impact on the achievement of students in their classes (like the moderation and standardisation at GCSE). (*English Teacher)*

> And we've had moderation meetings; we were together with the other schools, teachers in other schools looked at how rigorous our assessment would be and they criticised what, you know, our marking criteria is [*sic*]. And we changed it, which has all been very positive. (*Mathematics teacher)*

Another positive feature was illustrated by the following reflective statements from two teachers:

> I think the strength is that it's genuine. It's much more… it's much better for mathematics, I think much better for life. How have you thought about this? What's your solution to this? How else could you have done it? What other angles would you consider? What were the multiple answers you got? Rather than, "You got the right answer". […] This is much more satisfying. (*Mathematics teacher*)

> The kids felt it was the best piece of work they'd done. And a lot of teachers said, "OK we're really pleased with how they've come out". It's seemed to really give kids the opportunity to do the best that they could have done. […] The lower end kids—it organises their work for them; it's basically a structured path for them to follow. (*English teacher)*

These statements illustrate the general finding that students could be positively involved in tasks where they knew that the outcomes would form part of a summative assessment of their capabilities. Because the broader focus of their assessment work gave more opportunity for pupils to perform, the results revealed more to their teachers about on-going possibilities of improving engagement and motivation.

> You could see quite a lot from what people do, from how much work they put in outside of the classroom. […] And you can see quite a lot about how they think, as well. And how they work in groups is quite interesting. (*Mathematics teacher)*

Such effects were linked to the fact that teachers, being involved in the formulation of the tasks and the operational procedures for their use, had confidence in implementing them, so breaking down the gaps, in ownership and agency, between the formative and summative aspects of their work:

> But I think if all the teachers had more, … possibly more ownership of what we are actually doing in terms of summative assessment, then you would have more confidence in saying to parents, which I think is one of the biggest things I find with lower school. (*Mathematics teacher)*

Another teacher described a further outcome:

> I think it's quite a healthy thing for a department to be doing because I think it will encourage people to have conversations and it's about teaching and learning. […] It really provides a discussion, hopefully, as well to talk about quality and you know what you think of was a success in English. Still really fundamental conversations. (*English teacher*)

Such evidence, albeit from exploratory work in only three schools and in only two subjects, does show that development of teachers' competence with respect to summative assessment can have wider impact on their work. One reason for this is that if teachers are made responsible for producing summative assessments that are dependable and comparable between schools, they have to meet together regularly, to discuss aims, procedures and evidence, meetings which produce those sort of collegiate discussions which can be a valuable part of teachers' learning (Stiggins, Arter, Chappuis, & Chappuis, 2004). Teachers in one school, having experienced the benefit of the end-of-year moderation meetings, planned to have such meetings three times a year; they wished to explore the value of meetings to 'moderate formatively', i.e. to explore on-going progress in developing shared judgments and criteria in advance of the occasions where decisions would have to be taken.

However, this work took about two-and-a-half years during which the teachers had impetus and support from university researchers and advisory staff from their local authority. Teachers were also released from school for nine one-day meetings of the group as a whole. The extent to which such development can be sustained without such intensive support remains to be seen. A more detailed account of teachers' participation in moderation meetings, in the Australian state of Queensland where school-based assessments are well-established, having replaced external testing in the 1980s, has been published by Wyatt-Smith, Klenowski, and Gunn. (2010).

The links between formative and summative assessment

The evidence and arguments of this section serve two of the main aims of this paper, One is to demonstrate that teachers are able to contribute to the summative purposes of assessment in such a way that a system primarily designed, and used, to serve summative purposes, can in fact, also be supportive of learning. The other is to replace assumptions that a sharp discontinuity between the formative and the summative is the norm with acceptance that there ought to—and can—be synergy between them.

The discussion of this synergy in the previous section can be developed in more detail in two different ways. For the first, the work reported in Black et al. (2011) has shown how supportive links can be forged between formal summative assessments and the formative and informal summative functions that form part of box 5 in Figure 1. As teachers develop experience of implementing summative work in the ways that are needed to enhance the quality of their summative judgments, they will have to confront the need to produce better assessment instruments, so linking the design on teaching work with features (a) and (b) above, and to consider more deeply the validity of their judgments, so linking features (c) and (d) to this design and its justification in the light of the broad aims of their pedagogy.

A second way is to look more directly at the five key strategies of formative assessment, as set out in Figure 2 above. The several possibilities here were realised in the work of the KOSAP project.

The first and second of these five strategies in figure 2 are, of course, essential to any assessment designed to function summatively. The third may not be seen as essential for any summative assessment that is externally set and terminal, but where such assessment functions as an interim review which might indicate the need for repetitions of earlier work, or changes in the work of the next stage, it does serve as 'feedback that moves learners forward'. The last two relate to the ways in which students were prepared to engage in the production of work that would be assessed for summative purposes. For example, for work on tasks which require selection and use of mathematical skills in tackling a problem with an everyday context, students could work collaboratively to develop their understanding and experience of such work in ways that involve all of the strategies, and then be set a different task making similar demands which they have to tackle on their own. For a task requiring a report on a topic, which calls for research into a range of resources, pupils might share the work of assembling and analysing ideas and information, and then have to produce their

individual syntheses of these under controlled conditions. Such procedures were envisaged in the 'controlled conditions' rules for GCSE coursework in many subjects, which were implemented in England in the 2010-2011 school year (Qualifications and Curriculum Development Agency, 2009), but which were abandoned in 2015 because inadequate investment in procedures to ensure oversight of the implementation of the practices by individual teachers.

6 Formative and Summative Assessment in Pedagogy

One of the key ideas that has motivated this paper is that assessment is an essential aspect of effective education because the relationship between the instruction students receive and what they learn as a result of that instruction is complex. If students learned what they were taught, then assessment would be unnecessary; we could simply document the educational experiences of each student secure in the knowledge that this would describe their capabilities accurately. But of course, students do not always learn what they are taught, so we need to develop processes of eliciting and interpreting evidence so that we can draw conclusions about what students have in fact learned. From this simple view, there should be no conflict between formative and summative assessment—indeed, the distinction would not be useful—because all assessment would be about producing valid inferences about students.

A principle argument of this paper, however, is that assessment cannot be understood without a consideration of the wider context within which that assessment takes place. Teachers and schools are constrained, at least in the short term, by the cultural traditions, the political and public expectations of education, and the norms of the various institutions within which they operate. The ideas presented in this paper may illuminate for those involved in education some aspects of their working context, and suggest ways their practice might be developed. The ideas might also motivate changes, in which case consideration of the outcomes desired, of the tools available (or to be constructed), and of the feasibility of change within the rules, the accepted divisions of labour, and of the community, would merit examination. This, of course, is the language of activity theory, which is clearly relevant, but it is also complex in that there is not a single system, but a complex web of nested and overlapping systems, one with the school as subject, one with pupils or parents as subjects, one with governments as subject, and so on. What we hope is that the much simplified model presented in figure 1 provides a way of focusing analysis of the role of assessment in improving education by drawing attention to the key relationships and forces.

7 Conclusion

What is clear is that the features explored in the sections *3, 4, 5 and 6* above can only be understood, whether as stable or undergoing change, within the cultural, social and political situation within which they are embedded. In particular, one feature of pedagogy which these features do not address directly is Bernstein's concept of linguistic code that regulates cognitive orientation and moreover "regulates dispositions, identities, and practices, as these are formed in official and local pedagogizing agencies (school and family)" (1990 p.3), which implies that one of the main causes of educational success or failure is to be found in

the home. Thus, children from working class families, who are only familiar with the *restricted code* of their everyday language, may find it difficult to engage with the *elaborated code* that is required by the learning discourse of the classroom and which those from middle class families experience in their home lives. However, Bernstein does not explore the implications of this for classroom work: in the latest of his four books on "Class, Codes and Control' which has the specific subtitle *"Volume 1V The Structuring of Pedagogic Discourse"* the term 'assessment' is used only once where he draws attention to *'assessment procedures which itemize relative failure rather than the positive strength of the acquirer'* (p.87).

However, a link with the assessment issues in our paper is made clear in Robin Alexander's detailed analyses of his extensive data on classroom discourse, where he draws upon Bernstein's arguments. From the data that he has collected in many countries he highlights his findings that convergence of 'playground and classroom codes' is most evident in the English and American data and that it is "most obvious in group work" (p.481). Many features of our analyses also contribute to this convergence. Examples are the focus on interactive dialogue (pp. 7-9 and p.15), the emphasis on group work and collaborative learning (p.6), the argument that instrument used for summative purpose should also support learning (p.13), and the emphasis on students' active involvement (p.18). In all such work, students can be helped to transcend their lack of familiarity with elborated codes and teachers can thereby overcome the disadvantage suffered by some children because of the linguistic limitations of their family background

Thus our analysis cannot either give a full picture of assessment as it is conducted in any one context, or a deep understanding of why it is so conducted, let alone give advice about how change might best be pursued. What it can do is to give a framework for considering both the formative and summative aspects of teachers' assessments, which can illuminate the many confusions about the formative-summative relationship that lead policy and practice down blind-alleys, and also help the development of studies of pedagogy which might be better than most existing studies in giving due emphasis to the importance of assessment, linking it into the broader and complex overview which they seek to achieve.

References

Alexander, R. (2000) *Culture and pedagogy: international comparisons in primary education.* Oxford: Blackwell.

Alexander, R. (2004). Still no pedagogy? Principle, pragmatism and compliance in primary education. *Cambridge Journal of Education, 34*(1), 7-33.

Alexander, R. (2008) *Essays on Pedagogy*. London: Routledge.

Allal, L. (1988). Vers un élargissement de la pédagogie de maîtrise: processus de régulation interactive, rétroactive et proactive. In M. Huberman (Ed.), *Maîtriser les processus d'apprentissage: Fondements et perspectives de la pédagogie de maîtrise* (pp. 86-126). Paris, France: Delachaux & Niestlé.

Applebee, A.N., Langer, J.A., Nystrand, M. & Gamoran, A. (2003) Discussion based approaches to developing understanding: classroom instruction and student performance in middle and high school English. *American Educational Research Journal,* **40** (3), 685-730.

Assessment Reform Group. (2006). *The role of teachers in the assessment of learning*. London, UK: University of London Institute of Education.
Baines, E., Blatchford, P. and Kutnick, P. (2009) *Promoting Effective Group Work in the Primary Classroom.* London: Routledge.
Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles Policy and Practice, 18*(1), 5-25.
Berliner, D. C. (1994). Expertise: The wonder of exemplary performances. In J. N. Mangieri & C. C. Block (Eds.), *Creating powerful thinking in teachers and students: Diverse perspectives* (pp. 161-186). Fort Worth, TX: Harcourt Brace College.
Bernstein, B. (1990) *Class, codes and control volume IV: The structuring of pedagogic discourse.* Abingdon UK: Routledge.
Black, P. (2016). The role of assessment in pedagogy–and why validity matters. In D. Wyse, L. Hayward & J. Pandya (Eds.), *Sage handbook of curriculum, pedagogy and assessment* (Vol. 2, pp. 725-739). Thousand Oaks, CA: Sage.
Black, P. & Atkin, M. (2014) The Central Role of Assessment in Pedagogy. Ch. 38 pp. 775-790 in Lederman, N.G. & Abell, S.K. (eds.) *Handbook on Research in Science Education Volume II.* Abingdon, U.K. : Routledge.
Black, P. & Wiliam, D. (2007) Large-scale Assessment Systems: Design principles drawn from international comparisons. *Measurement.* **5**(1) 1-53
Black, P. & Wiliam, D. (2009) Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, **21**(1), 5-31.
Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-74.
Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*(4), 451-469. doi: 10.1080/0969594X.2011.557020
Black, P., Harrison, C., Hodgen, J., Marshall, M. & Serret, N. (2010) Validity in teachers' summative assessments. *Assessment in Education* 17(2) 215-232.
Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D, (2003) *Assessment for Learning–putting it into practice*. Buckingham: Open University Press.
Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2002). *Working inside the black box: assessment for learning in the classroom.* London, UK: King's College London.
Bruner, J. (1966) *Toward a Theory of Instruction*. New York: Norton for Harvard University Press.
Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2 ed., pp. 443-507). Washington DC: American Council on Education.
Daugherty, R. (2010) Summative assessment: the role of teachers. In Peterson, P., Baker, B. & McGaw, B. (eds.) *International Encyclopaedia of Education.* Volume 3, 348-391. Oxford: Elsevier.
Department for Education (2010a) *How can I use the three-part lesson?* London : UK Department for Education. Available on: nationalstrategies.standards.dcsf.gov.uk/node/85242 (accessed March 2011).
Dweck, C. S. (2000). *Self-theories: their role in motivation, personality and development*. Philadelphia, PA: Psychology Press.
Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin.

European Commission Directorate-General for Education and Culture. (2010). *Assessment of key competences: Draft background paper for the Belgian Presidency meeting for Directors-General for school education*. Brussels, Belgium: European Commission Directorate-General for Education and Culture.

Foley, B. & Goldstein, H. (2011) *League Tables in the Public Sector: A review and analysis.* London: British Academy Policy Centre

Freire, P. (1970). *Pedagogy of the oppressed* (M. Bergman, Trans.). New York, NY: Continuum.

Gardner, J. (2007). Is teaching a 'partial' profession? *Make the Grade: Journal of the Chartered Institute of Educational Assessors, 2*(Summer), 18-21.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*(4), 427-438.

Hallam, S. & Ireson, J. (1999) Pedagogy in the Secondary School. Ch.4 pp.68-97 in Mortimore, P. (ed.) (1999) *Understanding pedagogy and its impact on learning.* London: Paul Chapman.

Harlen, W. (Ed.). (2004). *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*: EPPI-Centre, Institute of Education, University of London.

Hayward, L. (2015) Assessment is learning: the preposition vanishes. *Assessment in Education: Principles, Policy and Practice,* 22(1), 27-343

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38*(5), 365–379.

Johnson, S. (2011). *A focus on teacher assessment reliability in GCSE and GCE*. Coventry, UK: **Office of Qualifications and Examination Regulation (OFQUAL)**

Kawalkar, A., & Vijapurkar, J. (2013). Scaffolding science talk: The role of teachers' questions in the inquiry classroom. *International Journal of Science Education, 35*(12), 2004-2027. doi: 10.1080/09500693.2011.604684

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.

Koretz, D. M. (1998). Large-scale portfolio assessments in the US: evidence pertaining to the quality of measurement. *Assessment in Education: Principles, Policy and Practice, 5*(3), 309-334.

Leach, J. & Scott, P. (2005) Perspectives on designing and evaluating science teaching. Ch.XX, pp.171-187 in Tomlinson, P., Dockrell, J. and Winne, P. (eds.) *Pedagogy – Teaching for Learning.* Leicester: The British Psychological Society.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 20*(1), 37-66.

McGuiness, C. (2005) Teaching thinking: theory and practice, pp.107-126 in Tomlinson, P., Dockrell, J. and Winne, P. (eds.) *Pedagogy – Teaching for Learning.* Leicester: The British Psychological Society.

Mercer, N., Dawes, L., Wegerif, R. and Sams, C. (2004) 'Reasoning as a scientist: ways of helping children to use language to learn science'. *British Educational Research Journal,* 30 (3): 359-377

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3 ed., pp. 13-103). Washington, DC: American Council on Education/Macmillan.

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology*. Menlo Park, CA: SRI International.

Moore, A. (2000) *Teaching and learning: pedagogy, curriculum and learning.* London: Routledge-Falmer

Mortimore, P. (ed.) (1999) *Understanding pedagogy and its impact on learning.* London: Paul Chapman.

NCETM (2007) *Was there ever any point to the three-part lesson?* London: National Centre for Excellence in Teaching mathematics. Available on: https://www.ncetm.org.uk/blogs/2401 (accessed March 2011)

New South Wales Board of Studies, Teaching and Educational Standards. (2012). Assessment for, of and as learning. Retrieved July 1, 2016, from https://syllabus.bostes.nsw.edu.au/support-materials/assessment-for-as-and-of-learning/

Newton, P. (2012) Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*. 10, pp.1-29.

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

Nystrand, M., Gamoran, A., Kachur, R., & Prendergast, C. (1997) *Opening dialogue: Understanding the dynamics of learning and teaching in the English classroom.* New York: Teachers College Press.

Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning. Towards a wider conceptual field. *Assessment in Education: Principles Policy and Practice, 5*(1), 85-102.

Popham, W. J. (2005, December 12). Personal communication with Dylan Wiliam Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Qualifications and Curriculum Development Agency. (2009). *Changes to GCSEs and the introduction of controlled assessment for GCSEs*. London, UK: Qualifications and Curriculum Development Agency.

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

Sandelowski, M., Voils, C. I., Leeman, J., & Crandell, J. L. (2011). Mapping the mixed methods–mixed research synthesis terrain. *Journal of Mixed Methods Research, 6*(4), 317-331. doi: 10.1177/1558689811427913

Schum, D. A. (1987). *Evidence and inference for the intelligence analyst* (Vol. 1). Lanham, MD: University Press of America.

Shulman, L. S. (2005, 6-8 February). *The signature pedagogies of the professions of law, medicine, engineering, and the clergy: Potential lessons for the education of teachers*. Paper presented at the Teacher Education for Effective Teaching and Learning, Washington, DC. http://www.taylorprograms.com/images/Shulman_Signature_Pedagogies.pdf

Shulman, L.S. (1999) Knowledge and teaching: Foundation of the new reform. Ch. 5, pp.61-71 in Leach, J. and Moon, B. (eds.) *Learners and Pedagogy.* London: Chapman.

Simon, B. (1981). Why no pedagogy in England? In B. Simon & W. Taylor (Eds.), *Education in the eighties: The central issues* (pp. 124-145). London, UK: Batsford.
Slavin, R. E., Hurley, E. A., & Chamberlain, A. M. (2003). Cooperative learning and achievement. In W. M. Reynolds & G. J. Miller (Eds.), *Handbook of psychology volume 7: Educational psychology* (pp. 177-198). Hoboken, NJ: Wiley.
Smith, F., Hardman, F., Wall, K., & Mroz, M. (2004) Interactive whole class teaching in the National Literacy and Numeracy Strategies. *British Educational Research Journal*, 30(3), 395-412.
Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of teacher assessment: Evidence of what works best and issues for development*. Oxford, UK: Oxford University Centre for Educational Assessment.
Stenhouse, L. (1985). *Research as a basis for teaching*. London, UK: Heinemann.
Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534-539.
Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning: doing it right—using it well*. Portland, OR: Assessment Training Institute.
Taba, H. (1962). *Curriculum development: Theory and practice*. New York, NY: Harcourt Brace Jovanovich.
TeacherNet. (2010). Useful lesson plans. *Teaching Resources.* Retrieved September 28, 2016, from http://webarchive.nationalarchives.gov.uk/20080520150154/http://teachernet.gov.uk/teachingandlearning/resourcematerials/resources/
Tyler, R. W. (1949) *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press.
Watkins, P. & Mortimore, P. (1999) Pedagogy: What do we know? Ch.1, pps. 1-19 in Mortimore, P. (ed.) (1999) *Understanding pedagogy and its impact on learning.* London: Paul Chapman.
Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester Jr (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053-1098). Greenwich, CT: Information Age Publishing.
Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. In A. Luke, J. Green & G. Kelly (Eds.), *What counts as evidence in educational settings? Rethinking equity, diversity and reform in the 21st century* (Vol. 34, pp. 254-284). Washington, DC: American Educational Research Association.
Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree.
Wiliam, D. (2016). *Leadership for teacher learning: Creating a culture where all teachers improve so that all learners succeed*. West Palm Beach, FL: Learning Sciences International.
Wiliam, D., & Black, P. J. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537-548.
Wiliam, D., & Leahy, S. (2015). *Embedding formative assessment: Practical techniques for K-12 classrooms*. West Palm Beach, FL: Learning Sciences International.
Wiliam, D., & Thompson, M. (2007) 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.

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.

Wiske, M.S. (1999) What is teaching for understanding? Ch. 17, pp.230-246 in Leach, J. and Moon, B. (eds.) *Learners and Pedagogy.* London: Chapman.
www.teachernet.gov.uk/resources/ (accessed March 2011).

Wyatt-Smith, C., Klenowski,V. & Gunn, S. (2010) The centrality of teachers' judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice* 17(1) 59-75

Figure 1: Model for Assessment in Relation to Pedagogy

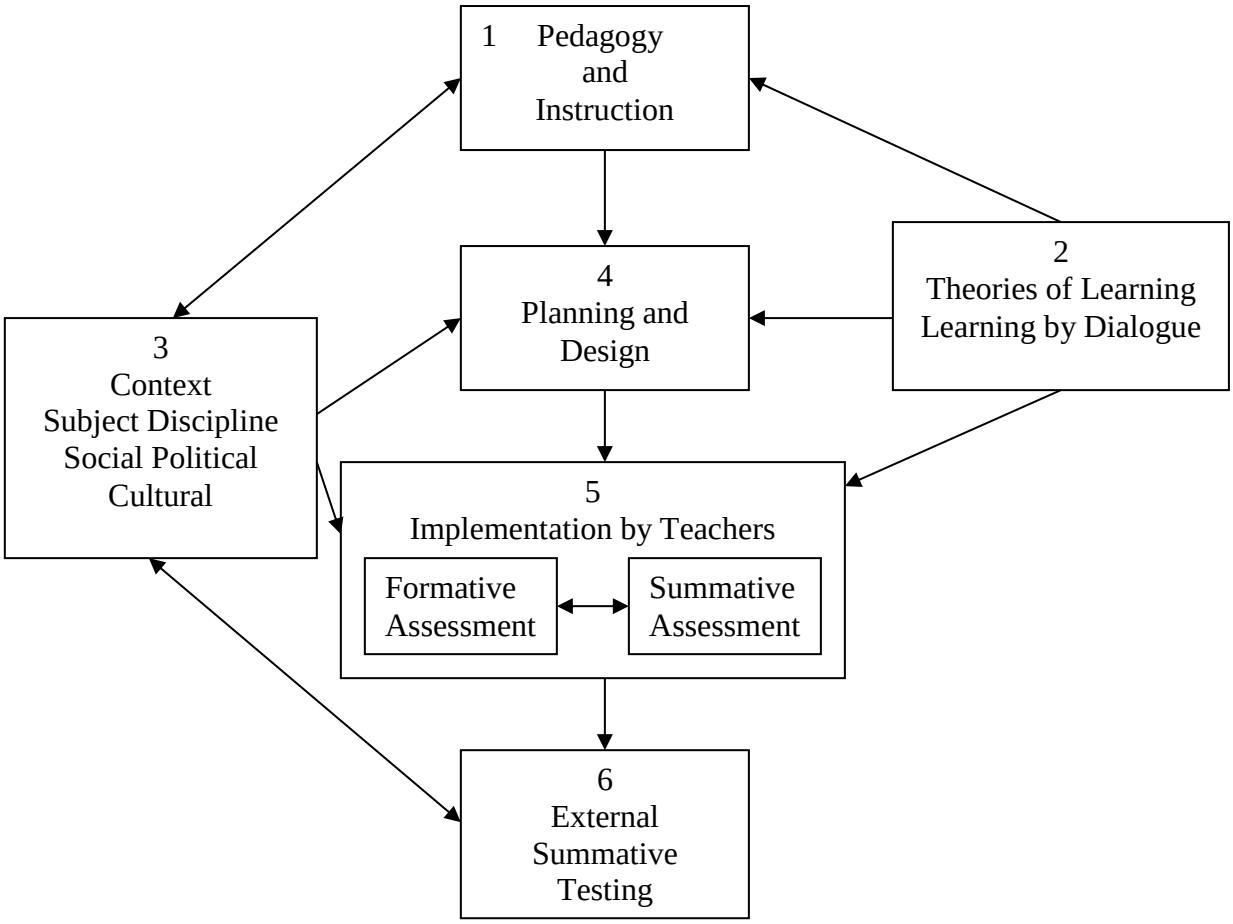

Figure 2: Key Strategies in Teacher Assessment

<table>
<tr><th></th><th>Where the learner is going</th><th>Where the learner is right now</th><th>How to get there</th></tr>
<tr><td>Teacher</td><td rowspan="3">1 Clarifying learning intentions and criteria for success</td><td>2 Engineering effective class-room discussions and other learning tasks that elicit evidence of student understanding</td><td>3 Providing feedback that moves learners forward</td></tr>
<tr><td>Peer</td><td Understanding and sharing learning intentions and criteria for success</td><td colspan="2">4 Activating students as instructional resources for one another</td></tr>
<tr><td>Learner</td><td colspan="2">Understanding learning intentions and criteria for success</td><td>5 Activating students as the owners of their own learning</td></tr>
</table>