Assessment in Education
Vol. 13, No. 3, November 2006, pp. 289–303

Routledge
Taylor & Francis Group

Some difficulties in informal assessment in mathematics

Anne Watson*
University of Oxford, UK

In this theoretical paper the informal assessment practices of two experienced teachers are used as cases for generating questions about future developments in formative assessment practice. Both teachers maintain a consistent formative assessment focus on the development of their students as enquirers, and one of them supplements this with explicit self-assessment activities. However, there are subject-specific gaps in the ways in which they assess and describe their students and these are not addressed in widely promulgated advice about formative assessment. Questions are raised about how teachers might be supported to develop their assessment of subject-specific behaviour.

Introduction

The recent development of national interest in the use of formative assessment to improve learning in the UK is an important example of research-driven change. Alongside the necessary evaluations of any such wide-scale implementation it is also useful to identify possible limitations of such approaches. In this paper, the work of two experienced mathematics teachers, who are aware of the powerful role of assessment in their teaching, is closely examined. These cases illustrate the existence of subject-specific gaps in the knowledge and practice of informal formative assessment. Suggestions are made about the nature of such gaps, and what might be necessary to bridge them.

I shall extract central findings from Black and Wiliam's research review (1998) to provide a framework for what follows, then proceed to report on the two teachers. Their work will be analysed by comparing their practice to the framework to see how far they fulfil the original research findings. Although their practices conform closely to recommended practice, gaps emerge which are not clearly identified except from a subject-content perspective, and which cannot be addressed by extant widely disseminated advice. Other avenues of potential support are explored.

*Department of Educational Studies, University of Oxford, 15 Norham Gardens, Oxford OX6 2PY, UK. Email: anne.watson@edstud.ox.ac.uk

ISSN 0969-594X (print)/ISSN 1465-329X (online)/06/030289–15

DOI: 10.1080/09695940601035445

Achievement and formative assessment

In their seminal study, Black and Wiliam (1998) found that innovative informal assessment practices, whatever their theoretical position, relate to significant learning gains. They identified five domains which make a difference to achievement, and I have summarized them very briefly here:

- *Goal orientation*: assessment should focus on processes rather than product, and should connect affective support and cognitive achievement. Praise is valuable as a reward for effort, but given too frequently, or in response to insignificant achievement, it can be demotivating (see also Pryor & Torrance, 1996).
- *Self-perception*: feedback needs to develop a personal sense of worth and progress; extrinsic rewards that focus on effort are more effective than those which focus on 'given ability' (see also Dweck, 1999).
- *Assessment by students*: giving students the responsibility for identifying gaps and how to work on them improves motivation and achievement (see also Schunk, 1996).
- *Links to theories of learning:* assessment is best seen as a moment in a holistic learning process, incorporating a reflective, meta-cognitive mode.
- *Teachers' strategies:* five areas of practice emerged as key to effective formative assessment:
  - *Choice of task*: tasks ought to relate to learning goals and provide opportunity to display evidence of achievement. Typical and unfamiliar tasks should both be used.
  - *Generation of discourse*: collaborative, exploratory, everyday forms are better than authoritarian and instructional discourse.
  - *Questioning*: higher-order questions, self-questioning, exploration of prior knowledge are more effective than procedural questions for formative purposes.
  - *Test use*: there is evidence that test results may not be used formatively.
  - *Feedback*: feedback about 'deep' learning and progression is important; it is more effective in complex, less structured tasks than in straightforward tasks; it should scaffold progress from what is already known.

Black and Wiliam's (1998) study was followed by their well-disseminated research project, King's-Medway-Oxfordshire Formative Assessment Project (KMOFAP), in which formative assessment practices were developed with teachers (Black *et al.*, 2002).

The research evidence for the effectiveness of formative assessment is robust, and KMOFAP succeeded in raising achievement by the equivalent of half a GCSE grade (Wiliam *et al.*, 2004). In addition, 4 comparisons between mathematics groups, involving 4 out of 12 mathematics teachers, yielded negative effects. Wiliam *et al.* (2004) provide plausible reasons for this, such as the varied nature of the comparison groups and the lack of experience of the teacher, but the question has to be asked whether there is something subject-specific which makes adoption of formative strategies inadequate to raise achievement on its own.

Informal formative assessment in practice

The two mathematics teachers whose work informs this paper see themselves as already doing many of the things advocated in the above summary; they have for many years included central features of the currently disseminated advice. Both teachers are highly experienced, and in many ways exceptional in their commitment and approach to the development of learner autonomy, yet both appear to have difficulty in focusing their assessment practices on subject learning. They were involved in separate research projects[1] each of which included a focus on classroom assessment of mathematics. Bob and Tania are not offered as typical, but examination of their work reveals the complexity of classroom assessment, seen from a subject-specific standpoint. Both teachers are working in the context of a highly specified curriculum with frequent national testing that is used to compare students, classes and schools. Both teachers are used to coping with administrative and political injunctions to 'do better', and actively seek out professional development support for the decisions they make in this climate. Both knew that their assessment practices were a focus for research, which was discussed with them at every stage.

In Bob's case, incidents in a lesson are analysed and used to illustrate assessment practices that relate closely to the research findings outlined above (Watson, 2000). Tania's case allows affirmation and further elaboration of complex practice, and includes additional, deliberate, practices resonant with those in Black *et al.* (2002).

An episode from Bob's teaching

Bob was teaching a series of lessons to a new-to-him Year 7 heterogeneous class in a UK comprehensive school. The class contained 30 students, who were from a largish market town which supported two other comprehensive schools and several private schools. The students in the school were mainly white, with English as their first language, and a significant number lived in nearby areas of rural deprivation. Achievement in the school was skewed towards a below national average level, and the proportion of socially underprivileged students was higher than the national average.

During eleven lessons, one each week for a term, I closely observed six students at work and tried to understand how Bob came to know about their mathematical learning. For this paper I focus on a student I shall call Garnet. The records I collected of Garnet's work include: all the written work which she handed in or did in her exercise book during the term; all the comments she made in whole class episodes during the eleven observed lessons; transcriptions of all overheard remarks and interactions, of whatever kind, made to peers or the teacher during those lessons; descriptive notes about her actions and behaviour. It was my intention in this research to have the same *type* of information as the teacher did, but inevitably to have a different collection due to my closer, but more intermittent, focus on Garnet and others. I did not test students, nor conduct interviews, because I wanted to learn more about the process of forming judgements based solely on what could be available to teachers in their normal practice. In addition I talked regularly with Bob about the six focus students to record the development of his thoughts and impressions of their work, and also, in the interests of equity, to give him access to my data to more fully inform the development of these impressions.[2] I report on episodes from a lesson, using observation and discussion notes as data.

Background of this lesson

In the second lesson that I observed, students were exploring properties of some 5-by-5 'magic' squares. Such objects are frequently used in UK mathematics lessons as an arena for exploring mathematical structure and number patterns. The idea is that numbers 1 to 25 are arranged in a 5-by-5 grid so that totals of each row and column and the two diagonals should be equal. Such squares have a number of mathematical properties that can be explored, generated, transformed and generalized by learners. These students had worked with simpler squares in earlier lessons and, when this incident begins, were verifying that sums of rows, columns and diagonals of some given arrays were equal. This lesson, the type of task set and the nature of Bob's interactions were typical of his teaching in all lessons I observed, in that students have to use familiar procedures to identify and manipulate multi-stage, complex and/or abstract, properties. They can then make general statements and justify them, and proceed to raise further questions of their own based on their findings. Rehearsal of techniques is embedded in the exploration.

Bob's belief about working with a new class at the start of secondary school was that he should be encouraging them to use exploratory, investigative methods in mathematics and to develop mathematical reasoning. His aim was to find out what each student was able to do, not in terms of mathematical knowledge but of fruitful mathematical behaviour. The extended task sequence about magic squares was intended to promote discussion, conjecture, comparison and reasoning, rather than any pre-specified content knowledge outcomes apart from revealing some of what they could already do. Thus it had both assessment and habituation purposes.

Observations of this episode

My first observation was that Garnet added some numbers using her fingers. She was adding 1 and 9 when she was interrupted by a friend and subsequently had to go back to the beginning. She was using a 'count all' strategy, counting up aloud to nine on fingers and then adding one more. Eventually, Garnet and Corinne (her neighbour), by counting everything on their fingers, muttering to each other, found that the row total for 5-by-5 squares should be 65.

At this point in the lesson Bob convened a whole class discussion to share the appearance of 65, which had arisen in several parts of the room at about the same time as it had for Garnet. At the end of this discussion he extended the task from 'verify that the sums are all equal to 65' to 'find sets of five different numbers which sum to

65'. (This is a typical task transformation in mathematics, where the result of exploring one question leads to the development of a new question.)

Working quietly together as a pair, Corinne pointed to some rows in the 5-by-5 squares she had already made and said to Garnet: 'Look, we can just use what we've already got, it's stupid, and when we run out we can use diagonals.' Garnet looked at her squares and muttered: '… use the diagonal'. For a while they rewrote the rows of their squares as horizontal addition sums. Eventually Bob approached and, not knowing that they had generated these horizontal sums from some magic squares, asked Garnet: 'How can you use these to make another magic square? That's the hard bit; it takes a bit of thinking.'

He reminded her about some earlier similar work with 3-by-3 magic squares. He pointed to some written work of Garnet's in which she had listed collections of three numbers which add up to 15; the collections had come from the rows and columns and diagonals of 3-by-3 magic squares.

Bob then left her to think how this related to her current problem. Garnet's answers in this interchange were quick, but later when I saw her written work I saw that the examples to which he had been pointing were not systematically constructed.

Towards the end of the lesson students were invited to contribute examples of sets which add to 65 and Garnet gave one, which was correct. She then asked in public: 'Are different arrangements of same numbers allowed?' Bob replied: 'Which choice gives fewer possible magic squares?' While others were paying attention to Bob's exposition, Garnet tried to rearrange a sum and use it to build a magic square. She muttered, out of his hearing: 'I don't know … no. No you can't have it'.

Analysing Garnet's mathematics in these incidents

Bob and I discussed these episodes after the lesson. My observations showed that Corinne had reasoned from A (the 5-by-5 magic squares) to B (the totals of 5 integers to make 65), having already established the property that the rows, columns and diagonals added to 65. Bob tried to get Garnet to see a way of getting from B to A without realizing that she and Corinne had generated B *from* A. Her responses were quick and correct but would have been hard to establish from the unsystematic way her bookwork was set out. This suggested to me that she already knew the connection from her memory and understanding of the previous work, whereas for him they confirmed that she now understood a connection he was helping her to make. He also knew that she had been able to contribute to the whole-class discussion, and had also generated an appropriate question about rearrangements of sums, and had responded positively to the challenge to answer it herself.

As an observer I was surprised by the 'count all' behaviour in a year 7 student, but learnt from Bob that weaknesses in arithmetic were not unusual and part of the purpose of the current task was to find out about such weaknesses in the context of exploratory activity. But the choice to use 'count all on fingers' does not necessarily indicate weakness, only that in this task she found counting on fingers an appropriate strategy, perhaps to keep track of *how* the total was accumulated.

From the lesson reported above the evidence suggests that Garnet may have thought about what she was doing and been aware of structure and properties, enough to make connections between different subtasks. In other words there was some metacognitive awareness. She also seemed to have mental images of 3-by-3 squares on which to call. She was also so keen to explore her own question that she opted out of listening to the teacher to do so.

A cumulative view of Garnet

After a few weeks Bob claimed that Garnet was 'quite aware'. He also claimed that she contributed a lot in class, although my notes did not confirm that she contributed any more than the average when I was there. When I drew this to his attention he said that because she sits next to Corinne, who definitely did have a lot to say, he often found himself looking in their direction. He said he *believed* that she was 'bright' and contributed frequently in whole class episodes. On the other hand, there had been a written test given in which students had been given a worked example of bracket use in algebra and asked to do a similar question. He said that Garnet had 'just guessed', incorrectly, and then sat doing nothing. This, he agreed, contradicted his belief about her engagement but he had dismissed the event as untypical.

I arranged to see her answers to a written test that had been recently administered to the whole cohort. It consisted mostly of arithmetic questions. In the test she had achieved 16 correct answers out of a possible 30, and Bob had told her she was above average for the group, the average mark being about 13.[3] If he had been concerned about arithmetic there was much to be learnt from the test, especially if we recall her adding 1 and 9 by 'counting all' on her fingers. He could, for example, have considered the possibility that she had poor knowledge of number bonds, or problems with place-value, but his beliefs about what was of value mathematically did not include these observations.

Over the term I observed many more interactions between Bob and Garnet. These were consistent with his stated belief that she could think things through for herself, and he questioned her, rather than telling her what to do. He seemed to be inducting her into self-directed thinking about, and participation in, mathematical tasks, rather than developing her specific knowledge as displayed in the test results.

To summarize Bob's assessment practices in relation to Garnet, as exemplified in the above episode and subsequent discussion:

- he observed her behaviour and interpreted it;
- he engaged in conversation with her using
  - his belief about the value of developing problem-solving strategies
  - his judgement that she needed his input
- he expected, based on past experience of similar groups of students working on this task, that she would be able to reason her own way through some features of the task, given some prompting.

This mixture of observation, interaction and judgment informed by belief, image and purpose is typical of teachers' informal assessment habits—indeed it could be used to describe many interpersonal human situations (Mavromattis, 1997; Watson, 2000). But this kind of assessment depends strongly on the teacher's aims, which could be about the learning of mathematical facts and skills, or mathematical structure, or how to behave like a mathematician, or some combination of these. In Bob's case, he consistently and coherently aimed at the development of mathematical behaviour in the nature of tasks, his interactions with students, the engagement expected of his students, and the focus of his informal judgements.

Tania's practice

Tania is an experienced teacher who was developing her work with a small class of students who were significantly underachieving in national tests before they entered secondary school. Like Bob's, they were mainly white, English-speaking, with a significant number from disadvantaged rural communities, bussed into a local market town which, in this case, had only one comprehensive school. Tania's work, like Bob's, took place within the context of a prescriptive curriculum, national expectations of lesson style, the provision of recommended teaching materials and comparison of schools in league tables based on test results. What follows is synthesized from notes and transcripts of interviews, discussions and lesson observations, and from Tania's own comments after discussion and analysis of these data. These data were collected during a year in which her teaching was observed and video-taped three times and she participated in six audio-taped discussion sessions with researchers and other teachers about her practice. In addition, she provided copies of planning documents and students' formal and informal written work throughout the year. The research project within which she was working was not primarily about assessment, but enough data about her assessment practices exists for the illustrative purpose I have here.

Tania's students had all previously failed to reach the government minimum target level, and she had decided not to use the officially provided teaching materials, nor to cover all the topics suggested in the officially provided schemes of work. Instead, she chose to structure her teaching to relate closely to what students already knew, with a focus on developing their mathematical thinking and understanding, while also ensuring that they learnt some facts and skills. In other words, her main purpose was similar to Bob's, but more clearly nested within the curriculum content of mathematics. Her plans were organized according to mathematical topics, rather than according to the development of exploratory learning habits. Within this framework she devised pre- and post-topic assessment tasks which would inform her not only about what they knew, but about their awareness of what they knew. She had two reasons for doing this. Firstly, that if she found out what they already knew, then she need not teach everything on the prescribed curriculum and could spend more time on fewer topics while still 'covering' what is required. There was more time available for consolidation activities, explorations and extended tasks. Secondly, there was a boost to learners' self-esteem if they were encouraged to express what they already knew at the start of a topic. She said: 'establishing what the students know helps to overcome the "I'm useless at maths" syndrome and allows students to move forward'.

She habitually used a range of methods for pre- and post-assessment. One example is the use of a 'sorting and describing' activity. For a statistics module, Tania gave each student an envelope containing small pieces of paper, each with a different word printed on them e.g. 'line graph, pie chart, frequency table'. They were also given glue and two sheets of paper. They were asked to stick on one sheet all the words they knew and on the other the words they did not know. On the first sheet they had to explain by writing and/or drawing what they thought each word meant. They could present the work in any way they wanted. Tania used these sheets to decide what she should teach and what she need not teach. This activity was given at the start and end of the topic to assess progress.

She also used other methods of focusing on content knowledge for pre- and post-assessment such as asking students to give examples of what they could do, to construct concept maps, and to create their own questions.

Towards the end of the year, Tania asked the students to comment in writing on their overall progress and attitude in mathematics lessons. She wanted to know if they felt positive about her lessons and were aware of improvements in their learning. This was the only self-assessment that was at all summative in function, and it was summative of their experience of her lessons rather than their specific mathematical knowledge.

Many claimed that they enjoyed mathematics more in Tania's classes and showed some insight into their own learning:

> I think I've made some progress towards maths because in my primary school I was always one of the ones who didn't understand, but now I'm in secondary school I seem to understand quite a lot of things in maths. (a student in Tania's class)

However, the value of this evaluation activity was doubtful. A few of the weakest in the group said that they were bored because the work was too easy and were perhaps unable to express the real nature of their problems. One student wrote, 'No I don't enjoy my maths lessons because there to easy and I end up not listen because I already Know them' (sic).

Yet on another occasion the same student had asked the researcher informally why she was videoing a lesson. The researcher had explained that she wanted to see how their mathematics had changed. This same student said, voluntarily: 'I think it has changed. I put up my hand much more. You feel that you know more. I feel much more confident.'

Response to self-reporting tasks, like any other activity, is influenced in part by the situation, audience, mode of communication and expectations of a norm. I do not know enough about the school's assessment practices, or previous schools, to infer how their responses are the product of enculturation into a particular discourse of self-assessment form-filling which may or may not reflect *actual* learning, achievement or attitude.

Tania's assessment practices include more semi-formal practices, and may seem to be more content-focused than Bob's. She offers a method of self-audit, she then plans her teaching to focus on the difficulties and gaps which these reveal. They later do a fresh audit, which tells her something about what they have learnt. But when she talks about students she shows that she has developed a rich sense of how each of them behaves while doing mathematical tasks of an exploratory kind, and what kinds of self-image they may have in relation to mathematics. While saying that she wants detailed knowledge of what they have learnt, for summative assessment she asked questions which led only to answers about affective aspects, rather than following on from the style of content-focused prompt used in her pre- and post-tests.

Comparing Bob's and Tania's assessment practices

Gipps *et al.*, (1995) offered three characterizations of teachers as assessors: 'intuitives', who had 'tried-and-tested' personal methods and resisted anything imposed from outside; 'evidence gatherers', who went through the motions of collecting work, test results, observation notes and so on which seldom related to their teaching, and 'systematic planners' who integrated focused assessment activities within their normal planning and teaching processes. Torrance and Pryor's (1998) later study suggested that these approaches overlap, and some teachers even sustain apparently contradictory practices, possibly due to lack of confidence, vulnerability and management pressures (p. 73). In these distinctions, Bob might appear to be an intuitive and Tania a systematic planner, but there are more subtle similarities and differences which emerge when we look at the *type and use* of evidence together, whatever the overall approach.

Bob seemed to be assessing in order to influence learning behaviour directly, rather than to influence his teaching. At no time during the term did he engage with Garnet's specific problems with arithmetic, nor use them as part of his overall picture of her ability to work mathematically and learn mathematics, even after the test results.

Tania claims to assess in order to inform and evaluate her teaching, rather than their learning, yet wants to help students build positive images of themselves as mathematicians, engaging in mathematical activity, becoming aware of their own learning, while also making progress by external standards. Thus she was concerned to collect information about knowledge and skills in ways which generated awareness of learning. She used, alongside the activities described above, a similar kind of integrated assessment for learning to that which Bob practised. She constantly re-planned her teaching, asking 'what should I teach and how should I teach it?', incorporating tasks which she hoped would foster exploratory thinking skills into her teaching of core curriculum topics, where Bob focused solely on exploratory skills, at least for the duration of the study.

It is revealing to compare Bob and Tania's practices critically to those described by Black and Wiliam (1998) and summarized at the start of the paper.

G*oal orientation* in both cases is towards process rather than content, but Tania relates affective to cognitive concerns more explicitly than Bob. Both teachers have a fundamental belief that all their students are able to work in exploratory ways and hence they give *feedback* which accentuates positive achievement about decision-making, reasoning and sense-making and is necessarily concerned with deep learning and complex, semi-structured work. Tania uses *self-assessment* but not necessarily to focus her students' awareness on 'gaps'. Both see assessment as moments in holistic learning experiences; both focus on *metacognition*, in that Tania encourages self-assessment and Bob focuses on learning how to be a better learner. Both of them model and encourage *exploratory, collaborative discourse*.

Comparing their work to what Black and Wiliam (1998) say about *choice of tasks* is less straightforward, because the idea that tasks can be either typical or unfamiliar seems to be orientated towards preparing students to do test questions. Dealing mathematically with unfamiliar tasks is, for many teachers, universities and employers, the overriding aim of mathematics education. In Bob and Tania's case the tasks are necessarily unfamiliar, not to prepare students for tests or employment, but to encourage exploration.

When we turn to *questioning*, neither of them are using recall and low-order questions, but nor are they using higher-order questions in general. Instead they use procedural, affective and metacognitive questions which are generic rather than topic specific. In this respect they are demonstrating a characteristic identified by Ruthven (1994) as a common weakness: 'teaching interventions tend to take the form of directing students' approach to a task, rather than refining their conceptualisation of it' (p. 448).

While Ruthven uses this observation to claim that a closer match is needed between task, instruction, achievement and key objectives, Bob's practice already shows a *very* close match of these if we consider his own objectives, since the task is to explore mathematics independently. To some extent Bob is justified in taking this stance by studies which show that a focus on mathematical thinking and problem-solving might lead to better learning (e.g. Boaler, 1997; Senk & Thompson, 2003; Watson & De Geest, 2005) yet his scaffolding seems, in the incidents I observed of which the one described is fairly typical, to be only vaguely supporting Garnet in doing what she cannot yet do on her own. There is no overt discussion about mathematics, or about mathematical thinking, in the data, but plenty of evidence of him encouraging her towards independent decision-making and implicitly modelling the kinds of questions she might ask herself. Tania, similarly, avoids higher-order questions about mathematics itself and focuses on affective and procedural support, taking an active role in directing approaches to tasks and focusing on mathematical content. In scaffolding learners' effortful thinking, a master-apprentice model appears to be at work in Tania's lessons, but in the development of mathematical concepts the dialogue seems more equally balanced between teacher and student (Morgan & Wyatt-Smith, 2000).

The summary provided by Black and Wiliam (1998) seems to be assuming contexts in which the normal focus is on content coverage and mastery, where self-assessment, metacognitive prompting, collaborative discourse, and high-order questioning would alter classrooms significantly. But Bob and Tania are *already* concerned with learning more about learning and yet their discussions with students are not about complex engagement with mathematics and their formative assessment practices do not provide explicit information about mathematical progress. Indeed, without a background model of the development of knowledge and mathematical thinking processes, formative assessment practices cannot fulfil this role, partly because these aims can be sidelined by teachers' laudable focus on holistic development and autonomous working.

How could Bob's and Tania's practice be improved?

For teachers like Bob and Tania, studies which focus on metacognition, such as Broadfoot *et al*. (2002), and the formative assessment methods reported in Black *et al.* (2002) support their existing practice. For example, they use extended wait times, open questions, guessing and risk-taking. In addition, the tasks they set offer opportunities for exploration, discussion and discovery. Their oral feedback is elaborated and metacognitive. Shifting responsibility to learners for monitoring and understanding their own learning is already a central theme in their practice. In many ways, therefore, they are exceptional teachers. Neither Bob nor Tania are primarily concerned with teaching students to pass tests, they both believe that becoming better learners of mathematics is an appropriate aim for students and that better test results are a product of that, rather than the other way round. Tania believes the sequence and pace of topics in the National Curriculum framework is inappropriate for her students, and is happy for them to learn any mathematics at all in any order; Bob believes the notion of order and hierarchy in mathematical topics is irrelevant for adolescents. Thus a hierarchical view of progress in mathematical knowledge offers neither of them any help because they do not subscribe to it, yet there are no easily available alternative descriptions which could map cognitive development in the subject. An appropriate direction for development of their practice would be to extend their questioning and task use so that it is more focused on the development of conceptual understanding, but not within the hierarchies they reject.

Large-scale testing often presents growth of mathematical knowledge as chronologically ordered or hierarchical, rather than the networked result of the sense-making constructions of adolescents. For these teachers like Bob and Tania it is not obvious how such detail might be used. Bob, you will recall, does not use the test results formatively and does not incorporate them into his view of Garnet's mathematics, indeed he does not seem to know what to do with them at all. Tania, while focusing on content, does not know how her students can assess their own content knowledge. But development of practice must start from what they already do, and injunctions to 'assess for learning', which they believe they already do, or to pay attention to national test results, are unlikely to lead to changes of practice unless they can do it in ways which do not compromise beliefs. Their natural focus for informal assessment is the development of mathematical habits. Testing, on the other hand, focuses on performance, not potential; on behaviour, not development. Yet the criteria used in testing, such as national standards or outcomes, provide an external frame of reference against which they could monitor their students' learning in ways which do not depend on local norms derived from in-class comparisons, or accumulated experience of past classes.

This is also a problem that influences the efficacy of self-assessment. If learners are to self-assess they need to understand what it is they can do (Sadler, 1989). When students self-assess *within the dominant discourse* of Bob and Tania's classrooms, they will be self-assessing their metacognitive skills, rather than their mathematics, or even their mathematical behaviour. There are no models in these classrooms of discussion about mathematical concepts on which students can model their self-assessment. A complex but comprehensible language connecting mathematical achievement to aspects of mathematical behaviour would be needed for these teachers to use individual target-setting and self-assessment effectively. We also need a substantial language about what it means to understand different kinds of mathematical topic—we do not have that. What we have instead are hierarchical curricula and testing requirements.

There are alternatives. A probabilistic spread of clustered understandings that does not assume hierarchy but accepts that individual learners construct their understandings in different ways would be helpful. Models are available for some mathematical topic areas. These claim to show how learners who understand topic X are likely to understand other related topics. Those provided by Callingham and others (Callingham & McIntosh, 2001, 2002) offer a model for teaching and learning mathematics that *fits* Bob and Tania's holistic aims, rather than detracts from it (Hattie & Jager, 1998, p. 121). Thus, while it is likely that a student who understands X will also understand Y, it is not assumed that this is an essential or hierarchical link.

It is important to draw distinctions between what a learner can understand from assessment criteria and can take responsibility for improving, and what has to be noticed by a teacher or requires teacher intervention. A teacher of a class of 30 students or more is unlikely to be able to spot students' habits since seeing patterns of behaviour requires sustained attention over time. However, this is precisely where national test feedback information that includes error analysis may be of help, since it could inform teachers about habits worth looking for in their students. In the test, Garnet failed to subtract successfully by decomposition. If Bob had had information to hand about the distribution of misapplied habits, such as taking smaller from larger, in his own classes he could identify what she had tried to do and devise appropriate tasks which expose and challenge her understanding. Furthermore, if teachers know that their students are likely to adopt specified unhelpful habits they can reconsider how they teach a topic and devise exploratory tasks which challenge precisely that tendency. Planning teaching according to common 'misconceptions' is a fairly common practice in the UK thanks to the informative power of the Concepts in Secondary Mathematics and Science study (Hart, 1981) but having 'smart' information about the likely conceptual development of *one's own students* is now a theoretical possibility through national test feedback, and seems a useful idea (Ryan *et al.*, 1998; Doig, this issue). However, if it is provided as a form of accountability, or as a comparative tool, it is unlikely that teachers will be able or willing to take time to extract and act on such specific data.

To be more radical, one might be tempted to say that Bob's and Tania's inclinations to focus on the development of mathematical thinking are right and it is progress in this domain that should be assessed. There have been several attempts to describe authentic mathematical activity, such as mathematical thinking and enquiry, as rubrics for assessment in order to 'measure' students' attainment as mathematicians (e.g. in the UK, and in Victoria, Australia). The result of such attempts is often to reduce such activity to a list of behaviours, so that the original purpose and awareness is lost. There is a great difference between someone who decides, in the middle of working on some mathematics, to generate a few special examples in order genuinely to find out what is going on, and someone who generates a 'systematic' set of examples because the teacher has told them they need to do this to gain marks. The latter is hardly worth doing. Authentic processes are always situated in real enquiry and exploration. They are fluent, intermittent, task-specific, and applied in unfinished contexts, and subject to the task-immersed perception of the learner. Otherwise they are inauthentic. For these reasons, it is hard to see how Bob and Tania could be any more explicit about the ways they would like learners to work on mathematics, without at the same time reducing such work to the fulfilment of certain behavioural objectives.

Emerging questions

This examination of classroom assessment from the point of view of the existing practices of two teachers who are experienced in developing holistic mathematical activity has highlighted gaps in practice and gaps in knowledge of classroom assessment. A series of questions emerges:

Can ways be found to use performance data from large-scale studies to construct relevant information for individual teachers?
Can non-linear pathways of mathematical development be described?
How can such descriptions be used by teachers and students without reducing mathematical enquiry to a rubric without purpose?

Acknowledgements

Grateful thanks are due to Judy Mousley, Helen Forgasz, Helen Doerr and Deakin University, Australia, who funded the preparation of this paper through a Qualitative Learning Research Priority Initiative. Frequent discussions with Rosemary Callingham and Brian Doig took place during every stage of the writing of this paper, and it is in every sense a product of those discussions.

Notes

1. I am disguising details of schools and students and use pseudonyms for the teachers and students, possibly changing genders, since some aspects of this paper could be taken to be criticisms of their work. Also, pseudonyms used in this paper do not match pseudonyms I have

used elsewhere. Further details of the projects of which these findings were a part can be obtained from the author.

2. At the time there was no expectation in UK research that Garnet and other focus students needed to know she was being especially observed, and it was deemed ethical not to tell them but to ensure teachers had access to all the data so that their informal assessments would be enriched and better informed than otherwise might be the case.
3. The test had been set for the whole year group, and involved four operations on whole numbers up to three digits, decimals with up to two decimal places, and fractions with simple denominators, as well as some early algebra.

Notes on contributor

Anne Watson is reader in mathematics education at the Department of Educational Studies, University of Oxford. She has many years' teaching experience in comprehensive schools, and was a member of a group within the Association of Teachers of Mathematics, which pioneered 100% portfolio assessment in mathematics at 16+ in the years 1988–1994. Her D.Phil thesis developed a description of teachers' informal assessment practices, and her subsequent research has focused in part on the contribution of teachers' judgements to equity, or inequity, in mathematics teaching. She is the proud mother of a mathematics teacher whose students regularly outperform all institutional expectations. She publishes widely on classroom practice and also researches the relationships between interactive practices and the development of mathematical thinking.

References

Black, P. & Wiliam, D. (1998) Assessment and classroom learning, *Assessment in Education: principles, policy and practice,* 5(1), 7–73.

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2002) *Working inside the black box: assessment for learning in the classroom* (London, King's College London).

Boaler, J. (1997) *Experiencing school mathematics: teaching styles, sex and setting* (Buckingham, Open University Press).

Broadfoot, P., Daugherty, R., Gardner, J., Harlen, W., James, M. & Stobart, G. (2002) *Assessment for learning: 10 principles.* Assessment Reform Group. Available at: www.assessment-reform-group.org.uk/principles.html. (accessed 8 August 2003).

Callingham, R. & McIntosh, A. (2001) A developmental model of mental computation, in: M. Mitchelmore, B. Perry & J. Bobis (Eds) *Numeracy and beyond: proceedings of the 24th Annual Conference of the Mathematics Education Research Group of Australasia,* (Sydney, MERGA), 130–138.

Callingham, R. & McIntosh, A. (2002) Mental computation competence across years 3 to 10, in B. Barton, K. Irwin, M. Pfannkuch & M. Thomas, (Eds) *Mathematics education in the South Pacific: proceedings of the 25th Annual Conference of the Mathematics Education Research Group of Australasia,* (Sydney, MERGA), 155–163.

Doig, B. (2006) Large-scale mathematics assessment: looking globally to act locally, *Assessment in Education: principles, policy and practice,* this issue.

Dweck, C. (1999) *Self-theories: their role in motivation, personality and development,* (Philadelphia, Psychology Press).

Gipps, C., Brown, M., McCallum, B., & McAlister, S. (1995) *Intuition or evidence? Teachers and national assessment of seven year olds* (Buckingham, Open University Press).

Hart, K. (Ed.) (1981) *Children's understanding of mathematics: 11–16,* (London, John Murray).
Hattie, J. & Jaeger, R. (1998) Assessment and classroom learning: a deductive approach, *Assessment in Education: principles, policy and practice,* 5(1), 111–122.
Mavrommatis, Y. (1997) Understanding assessment in the classroom: phases of the assessment process—the assessment episode, *Assessment in Education: principles, policy and practice,* 4(3), 381–399.
Morgan, W. & Wyatt-Smith, C. (2000) Im/proper accountability: towards a theory of critical literacy and assessment, *Assessment in Education: principles, policy and practice,* 7(1), 123–141.
Pryor, J. & Torrance, H. (1996) Teacher-pupil interaction in formative assessment: assessing the work or protecting the child?, *The Curriculum Journal,* 7(2), 205–226.
Ruthven, K. (1994) Better judgement: rethinking assessment in mathematics education, *Educational Studies in Mathematics,* 27(4), 433–450.
Ryan, J., Williams, J. & Doig, B. (1998) National tests: educating teachers about their children's mathematical thinking, in A. Olivier & K. Newstead (Eds) *Proceedings of the 22nd Conference of the International Group for the Psychology of Mathematics Education,* (South Africa, Stellenbosch), 4, 81–88.
Sadler, R. (1989) Formative assessment and the design of instructional systems, *Instructional Science,* 18(2), 119–144.
Schunk, D. (1996) Goal and self-evaluative influences during children's cognitive skill learning, *American Education Research Journal,* 33(2), 359–382.
Senk, D. & Thompson, D. (2003) *Standards-based School Mathematics Curricula: what are they? What do students learn?* (Mahwah, NJ, Erlbaum).
Torrance, H. & J. Pryor (1998) *Investigating formative assessment: teaching, learning and assessment in the classroom* (Buckingham, Open University Press).
Watson, A. (2000) Mathematics teachers as informal assessors: practices, problems and recommendations, *Educational Studies in Mathematics,* 41(1), 69–91.
Watson, A. & de Geest, E. (2005) Principled teaching for deep progress: improving mathematical learning beyond methods and materials, *Educational Studies in Mathematics,* 58(2), 209–234.
Wiliam, D., Lee, C., Harrison, C. & Black, P. (2004) Teachers developing assessment for learning: impact on student achievement, *Assessment in Education: principles, policy and practice,* 11(1), 49–65.