TEACHER ASSESSMENT OF SCIENCE IN ENGLISH PRIMARY SCHOOLS

Dan Davies[1], Sarah Earle[2], Christopher Collier[2], Rebecca Digby[2], Alan Howe[2] and Kendra McMahon[2]
[1] Cardiff Metropolitan University
[2] Centre for Research in Early Science Education (CRESL), Bath Spa University

Abstract:

Teacher assessment of pupils' on-going classroom science work can be a more valid means of judging their attainment than testing, because it can be based on a wider range of evidence; for example, observations, discussions and pupil presentations. However, questions remain regarding the reliability of teacher assessment, as they find such summative judgements difficult to make and have limited opportunities for moderation. This is of particular concern in England, where assessment of the new National Curriculum in science is entirely dependent on teacher assessment at primary level. The Teacher Assessment in Primary Science (TAPS) project aims to develop the quality and reliability of teachers' judgements and establish a set of principles for effective assessment against which schools can evaluate their practice. It is driven by the following questions:

RQ1: What approaches are primary teachers in England currently using to assess pupils' learning in science?

RQ2: How valid, reliable and manageable are these approaches?

RQ3: Can an approach be synthesised from elements of existing practice which embodies core principles of effective assessment?

Data collected from 12 schools have included samples of assessment materials, interviews with key staff and observations of teachers making assessments during science lessons, together with school self-analyses against a theoretical framework developed from the work of Nuffield Foundation (2012), which uses the analogy of an ecosystem pyramid of numbers to represent the flow of assessment information between formative and summative purposes. Each participating teacher has also developed and piloted classroom assessment tasks, the resulting pupil work from which has been moderated between schools. Qualitative analysis of the above data sources has fed into case studies representing a typology of approaches examined against the principles of assessment embodied in the self-evaluation framework.

Keywords: teacher assessment, primary, formative, summative

INTRODUCTION

Science assessment at primary school level in a number of countries (e.g. Finland, Australia, New Zealand, Scotland, Northern Ireland) relies upon teachers' professional judgement of pupil attainment and progress, based on evidence gathered through everyday classroom work. Sahlberg (2011: 23) cites three main reasons why Finnish education relies on teacher assessment rather than external testing: 1. The progress of each pupil is judged more against his or her individual development and abilities rather than against statistical indicators; 2. assessment is embedded in the teaching and learning process and is used to improve both teachers' and pupils' work throughout the academic year; 3. academic performance and social development are seen as a responsibility of the school, not external assessors. Even countries formerly assessing science at age 11 through national tests (e.g. England, Wales) have shifted towards teacher assessment, reflecting a growing awareness of the harmful effects of high-

stakes summative testing (Newton 2009) together with its distorting effects on the taught curriculum (Wiliam 2003) and - in Wales at least - increased trust in teachers' professionalism (Daugherty 2009).

This increasing reliance on teacher assessment raises a number of issues. The first is the extent to which evidence of pupil learning collected for the formative purposes of supporting their further development can legitimately be used to summarise attainment against external criteria. Harlen (2013) asserts that any assessment opportunity can be used for both formative and summative purposes. The 'day-to-day, often informal, assessments' (Mansell et al. 2009, 9), which are used to inform next steps in learning, can also be summarised at a later date, whilst, conversely the results from summative tests can be used formatively to guide learning (Black et al 2003). However, Gipps and Murphy (1994: 14) argue that 'any attempt to use formative assessment for summative purposes will impair its formative role', though Wiliam and Black (1996) argue that this is possible as long as the elicitation of evidence is separated from its interpretation or judgement.

Validity, reliability, manageability and impact in teacher assessment

Other questions over the role of teacher assessment concern the extent to which it can meet key criteria for effectiveness, including validity, reliability and manageability. Each of these terms has a variety of meanings, so merit a brief discussion. Validity in assessment describes its intrinsic quality in capturing learning for the purposes to which the resulting data are to be put. Gardner et al. (2010) argue that teacher assessment has greater validity than testing because it can be based on a wider range of evidence. This is particularly relevant to a multi-faceted activity such as scientific enquiry, involving collaboration, the application of practical skills and problem-solving (Kelly and Stead 2013). However, Stobart (2012) asserts that validity is principally tied to the purpose of the assessment, which raises the question of whether validity is compromised if the same evidence is used for multiple purposes. Furthermore, there may be more than one type of validity which teacher assessment is required to meet. For example, content or construct validity concerns the extent to which a particular assessment instrument represents the range of skills and understanding for a particular topic area. Construct under-representation is a threat to validity, especially in primary science where the key skills of scientific inquiry may be more difficult to assess in classroom environments – though arguably more easily than in a test. Construct irrelevance is also a danger; Johnson (2013: 99) found teachers to be consciously or unconsciously influenced by construct-irrelevant pupil characteristics (e.g. gender, ethnicity, socioeconomic status). Other relevant forms of validity include participant-confirmed validity, whereby the pupil recognises themselves in the judgement of the teacher – highlighting the role of self and peer-assessment – and consequential validity, concerning the extent to which information gathered for formative purposes is used to support pupil learning. This form of validity could be termed impact, since a criterion for effectiveness in assessment should be the benefit it confers to learners, by contrast with the negative impact (in the form of stress, anxiety and repetition) of external testing.

The reliability of an assessment – broadly definable as the extent to which the results can be said to be of acceptable consistency or accuracy for a particular use (Harlen 2013: 9) - can also take a variety of forms, including inter-rater reliability, which concerns whether the same judgement would be made on the same evidence by different teachers. However, whilst teacher assessment has potentially greater validity than testing for the reasons outlined above, there remain concerns over its reliability (Black et al. 2011), particularly when it involves rating annotated samples of pupil work against external criteria (Klenowski and Wyatt-Smith 2010). External testing using standardised instruments can be argued to produce results of greater consistency, whose reliability is measurable. Whilst few studies have attempted to assign coefficients of reliability to teacher judgements, it is widely acknowledged in the

countries listed above that the most effective way to improve reliability of teacher assessment is through consensus moderation (Johnson 2013). Some jurisdictions such as Queensland also employ external moderation and exemplification of criteria to support teachers' judgements (Klenowski 2011), though studies of moderation processes have found that it takes up to three years to achieve acceptable inter-rater reliability through such approaches (Stanley et al (2009). Wiliam (2003) argues that, whilst teacher assessment can become more reliable, there is inevitably a 'trade off' between reliability and validity since the wider range of evidence required to represent the constructs within a field of learning may be more difficult to rate consistently on an external scale than the relatively narrow dataset obtained through a single assessment instrument.

This trade-off also relates to the manageability of teacher assessment, which requires that 'the resources required to provide an assessment ought to be commensurate with the value of the information for users of the data' (Harlen 2013: 10). Clearly a balance between and optimisation of validity, reliability and manageability is required for any effective approach to teacher assessment. However, teachers also need to develop a shared, secure understanding of assessment, particularly in a time of change in assessment policy (Brill and Twist 2013). If teachers do not have an explicit view of what constitutes effective assessment in science – which Klenowski (2011) has termed 'assessment literacy' - then capacity for improvement will be limited.

METHOD

Primary schools in the South West of England were invited to apply to participate in the Teacher Assessment in Primary Science (TAPS) project. We selected 12 schools from 50 applications to represent a mixture of large and small rural and urban settings, together with a range of approaches to teacher assessment. The project strategy is to intersperse central cluster meetings with visits to project schools, collecting examples of good practice in science assessment and facilitating school developments in assessment approaches and processes.

RQ1 was addressed through interrogation of a national UK database (The Primary Science Quality Mark, see Earle (2014)) and through collecting case study material from TAPS project schools. The data upon which this paper draws were gathered during five visits to each school between November 2013 and December 2014, involving interviews with science, assessment and ICT coordinators (n=36), observations of science lessons from Years 1 to 6 (n=72), collection of school science and assessment policies (n=24), collection of examples of assessment tools (n=36), annotated pupil work (n=144), tracking grids (n=12), formats for reporting to parents and pilots of focused assessment tasks linked to the 2014 National Curriculum (n=24). We used an analytical framework (see below) to evaluate qualitatively the validity, reliability, manageability and impact of each school's approach to teacher assessment, reporting findings as a series of case studies.

Towards an analytical framework for teacher assessment

The theoretical model of whole-school assessment upon which the TAPS project is based was proposed by a working group of experts convened by the Nuffield Foundation (2012) to consider the development of policy, principles and practice in primary school science assessment. This model considers the flow of assessment data through a school as analogous to the flow of energy through a biological ecosystem, with the various purposes for the assessment (formative, monitoring, summative, reporting) conceived as trophic levels within a 'pyramid of numbers'. Whilst at the formative, classroom level at the base of the pyramid a wide range of evidence would be used to inform teaching and feedback to learners, only a proportion of relevant data would feed up to the next level - which in the Nuffield model is

reporting to parents – and successively smaller selections of assessment information would flow upwards to the other summative levels of the pyramid.

To populate this model with criteria by which the quality of a school's science assessment might be evaluated we used the set of standards emerging from the Global Network of Science Academies conference *Developing Inquiry-Based Science Education: New Issues* held in Helsinki in 2012 (Harlen 2013). Through piloting these statements as self-evaluation criteria with the TAPS project participants through 2013-14 we allocated them to the levels of the Nuffield model, rearranging these levels and adding a further one corresponding to the role of pupils in formative assessment in response to feedback from the schools involved. From this emerged the analytical framework in Figure 1, which we have cross-referenced with the key criteria of validity, reliability, manageability and (positive) impact referred to above.

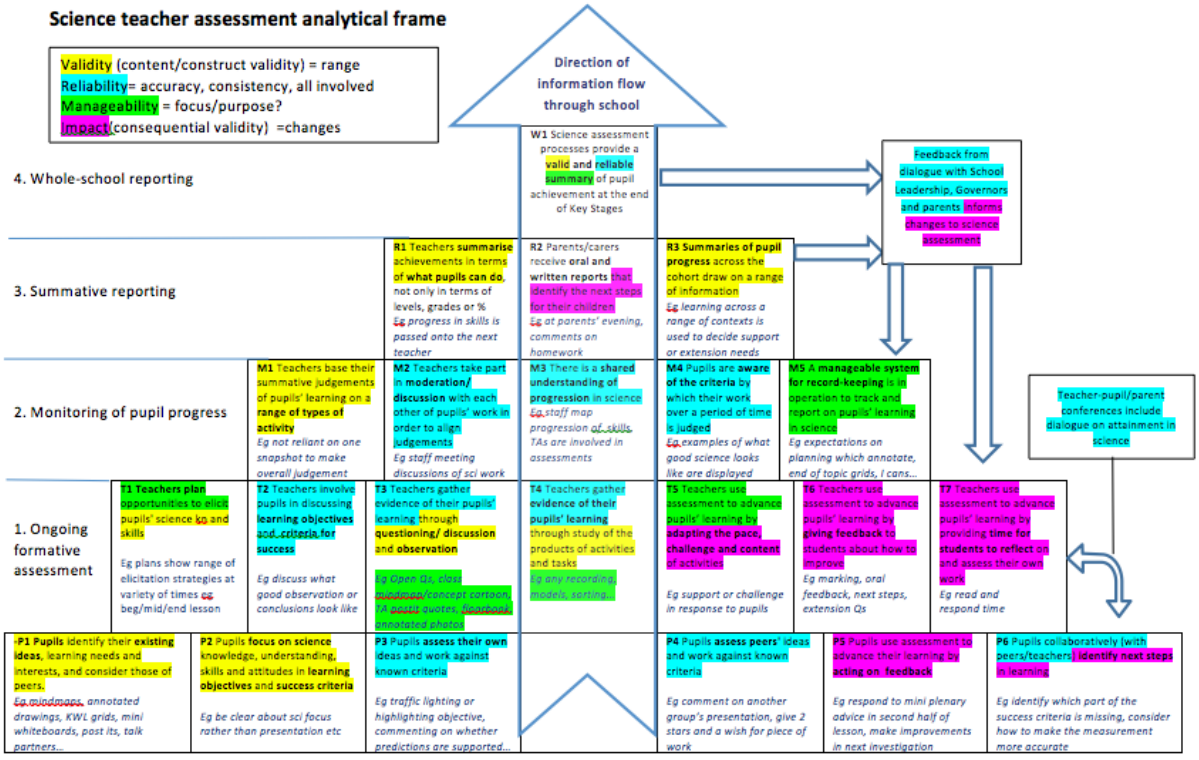

Figure 1: Science teacher assessment analytical frame

RESULTS

In order to answer RQ2, data from each case study school were evaluated against the criteria in the analytical framework. Each cell in the pyramid was allocated an alphanumeric code (e.g. S1 for the first cell in the summative reporting level, M3 for the third cell in the monitoring level etc.) with a summary data table completed for each school. The five case studies below illustrate the range of approaches taken and their relative strengths in validity, reliability, manageability and impact.

School A: strong validity in formative strategies with weak reliability in summative judgements

Pupil level

Pupils are involved in discussing learning goals through the collaborative process of constructing a 'Learning Wall' as a whole class (observed 11.13, 3.14 and 5.14) (P2). A 'Learning Wall' is a display board in the classroom that is used to document the development of a topic for the whole class, using pupils' drawing and writing and photographs, annotated by the teachers for younger pupils. Individuals or groups develop *KWL grids* (What do I

know? What do I want to know? What have I learned?) or *Mind maps* that identify relevant prior knowledge the pupils have and what questions they have about the topic (P1).

Teacher level

The above sharing of learning goals and expectation was also visible within observed lessons, with a range of devices being used to engage the pupils in discussing or considering these 'critical/key skills' - What do we need to be like? What will I see? What will I hear? – 'WALT' (What we Are Learning Today), use of exemplification – 'WAGOLL' (What a Good One Looks Like), 'Think - pair –share' (T2). Teachers gather evidence of pupils' ideas using a range of strategies, including photographed observations of their actions and notes of their utterances (lesson observation 11.13) (T3) . This breadth of strategies across the full range of the science curriculum provides high construct validity. Teachers adapt the pace and challenge of lessons in response to their assessment of pupils' learning (T5). For example, in a lesson on electric circuits with 5-6 year-olds (observation 11.13), the teacher had listened carefully to a pupil's suggestion that they use a lower voltage battery to see whether a buzzer would make a quieter sound, adapting her planned activity of switch-making to test this hypothesis. A sample of pupils' science exercise books from every year group shows that, across the school, teachers are providing pupils with feedback through their marking. Sometimes this takes the form of a simple question that requires an answer; sometimes multiple options are offered for the pupils to circle the idea they think is the best (T6). In every case the child had given some form of response to the feedback (P5). Science subject leader (SSL) monitoring of formative assessment indicates that its impact (consequential validity) is positive:

> I think, looking at the planning (T1) and book scrutiny, they're obviously getting formative assessment from the next steps being identified and pupils responding to that….I think that's pushing them and extending them enough. (SSL interview 29.11.13)

Monitoring level

School A's view is that summative record keeping should be minimal and formative assessment is where they should be focussing their energy. Rather than lengthy 'write ups' of practical work they make a careful choice of what pupils record in written or diagrammatic form – e.g. a drawing of a circuit that was made, a 'graffiti wall' of a group's ideas, a poster plan of a design for an ideal habitat (lessons observed 11.13-5.14) – from which teacher make judgments on their levels of understanding (M1). These summative 'best-fit' grades are recorded for each child three times a year (M5). The SSL has devised record sheets for each science topic that show three possible levels of outcome and the teacher writes the child's name in one of the three boxes. In the interests of manageability, teachers are not required to produce any further evidence of pupils' learning to support their judgments, as explained by the SSL:

> I think we're in danger of putting too much on teachers and then actually the fun of science goes … I want them to still be excited about it because that's what it should be all about. (SSL interview 29.11.13)

However, the SSL is considering how the reliability of teacher's judgments can be improved and demonstrated to an external audience. This may take the form of introducing moderation for science (M2) or focussed assessments of scientific inquiry skills, since there is evidence from her monitoring that these are under-represented in the range of assessment activity (M1).

Summative reporting and whole-school levels

Pupil achievement is discussed with parents in terms of what they can do, not only levels or grades (S1). For most year groups, pupil attitude towards science is an important focus for these reports, which do not signal 'next steps' in learning (S2).

A single numerical grade for science in the year is provided for each pupil and recorded on a central database for external reporting (S3). This represents an extreme reduction in data between the lower two levels of the pyramid model – where there is evidence of rich, valid assessment information having a positive impact on teaching and learning – and the whole-school level, where the extent to which these summary grades are informed by valid evidence and reliable judgements is unclear.

School B: a variety of formative strategies unconnected to summative assessment

Pupil level

At School B, science lessons (observed across five year groups during 2013-14 in a variety of topic areas including bones, rocks, electrical circuits and changes of state) often begin with a recap of learning from the previous session (P1). Pupils are targeted to take the lead with some 'revision' of previous learning because the teacher has noted they were unsure of a concept last lesson. '*Let's remind ourselves what a 'fair test' is.*' '*What do* factors *or* variables *mean?* Teachers then involve pupils in discussing learning objectives (P2, T2), taking care to ensure they understand the meaning of key words that will be used during the lesson. At the end, pupils are asked to choose a learning objective and say 'I can…' (P3). School B is exploring ways in which pupils can be given a more central role in gathering evidence of learning to aid reflection, inform assessment and facilitate better feedback.

Teacher level

Once the lessons are underway teachers gather evident of pupils learning using a range of strategies, such as p*artner 'buzz-time'* discussions, to respond to searching questions such as *what do batteries have inside them*? Questions are also with individuals or groups as practical work is undertaken (T3). Teachers will note where the pupils need to be reminded to focus on learning objectives, and intervene appropriately: *It's important to explain …why? Let's predict what is going to happen. What are you going to measure?* (T6). Opportunities for dialogue might be planned throughout the lesson. Pupils are happy to seek help: the child says *I'm getting confused*, so teacher explains in different way or clarifies the task. Questions are used to encourage pupils to recap on learning, e.g. *How did you do that?* (T7). There might be a 'mini-plenary' where a recap of the first activity occurs, in which scrutiny of pupils' group investigation planning sheets allow for a quick assessment of where each group has reached (T4). Teachers will check progress by reviewing pupils' writing, drawing, diagrams and charts and give written feedback (T6). They note which pupils have found the concepts difficult and will respond through marking and annotating books (T6), revisiting concepts at the next opportunity or amending planning (T1). A teaching assistant might offer feedback on a group's teamwork skills or an individual's progress against the learning objective. If a term is not understood, pupils are asked to dwell on its meaning: *Let's think about what 'durable' means. Do we mean* melt, *or* dissolve*?* (T5). The whole class might look at photos of the lesson in progress and review their learning (T7) or apply new knowledge learned during plenary sessions (e.g. *What is missing from my circuit*?). These approaches are making a positive impact on learning and teaching, demonstrating the consequential validity of the approach.

Monitoring, reporting and whole-school levels

During 2013-14 School B had no systematic approach to assessing or recording pupil attainment and progress in science. There was clearly a need to capture relevant evidence

from the formative processes described above to fulfil the analytical framework criteria for effective and reliable teacher assessment for summative purposes.

School C: Strong reliability in attainment tracking through evidencing and moderation

Pupil level

Pupils in School C make use of peer assessment to evaluate each other's learning against known criteria and inform next steps (P4, P6). The following examples are drawn from annotations on each other's work:

> I like your explanation about dissolving and your sentence about the liquid dissolving the powder. Improve: maybe label the other two jugs in the picture because I don't know what's in them. (Peer feedback, 11.13)
>
> I like K's technical vocabulary like pollination. The labels are clear and go in the correct order. (11.13)

Teacher level

During interview, (1.14) the assessment co-ordinator (AC) at School C identified a number of formative strategies in common use across the school:

> So you'll see speech bubbles in the books…. we also do written feedback (T6)… it might just be a question for them to answer, so they're sort of reflecting on it (T7)… make notes on post-its of observations of what the pupils have said (T3) and then there'll be photographs to go in and support it so that… when we come back to assessment, it's really clear.

One type of annotation on photos of pupils doing science is a record of what a particular child said in response to the teacher's question (T3), which provides important evidence for an assessment judgement, as in the following example:

> (T): 'Why did you choose these objects? (P): Because they're all flexible. (T): How do you know they're flexible? (P): Because they all bend.' (annotation on photograph of Y2 pupil's sorting of materials)

Another purpose for annotating examples of pupils' work is to enable teachers to provide formative feedback to support progression (T6) as below:

> Extension: check your predictions and mark if they are correct. Next step: to observe what changes happen when a material is heated. (teacher annotation on child's grid of predictions about how materials will behave when stretched, twisted, bent or squashed)

Monitoring level

To record and monitor pupils' progress in science (M5), School C uses individual pupil tracking grids on Excel spreadsheets developed by the UK government-sponsored Assessing Pupil Progress (APP) initiative (2010). To enhance the reliability of teachers' judgements against the criteria within the APP grid, the SSL set up a system in 2013-14 to cross-reference these judgements to evidence in the form of annotated samples of pupil work and dated observations, in order to address '… inconsistencies in judgements made using APP and what constitutes suitable evidence' (School Development Plan 2013-14). To support colleagues in generating such evidence, the SSL introduced a series of assessment booklets linked to particular criteria:

> … its got in it tasks that relate to the different units in each year group… they tend to very much either be around using their skills to design something new, or carrying out an experiment to find something out or explaining something, so it does test different skills, but then attached to it, you get the APP points that match that unit.' (SSL interview, 11.13)

There is evidence that these booklets support staff understanding of progression (M3) and their collection of evidence (M2):

> Sometimes it is hard to think of an investigation for some topic areas. So just getting that sort of starter and it worked well because we managed to then differentiate… We do it every term, so six times a year. (AC interview 1.14)

Annotation of the samples of pupil work linked to the tracking grid provides scope for additional contextual information – such as the degree of adult support provided – to be included, providing a degree of flexibility in the matching of evidence to particular APP criteria (AC interview 1.14). The reliability of this process is further enhanced through consensus moderation between groups of three colleagues, three times per year (M2). These 'triangulation' groups look at planning, science books and criterion-referenced judgements together, using a sub-sample of pupils's work drawn from three points in the attainment range. Teachers also have the opportunity to moderate with colleagues in other local primary schools, which the Assessment Co-ordinator believes has had a positive impact on reliability:

> I've noticed at moderation, it's more consistent. So that if you do move schools, you know that a 2c (criterion-referenced judgement) at that school is the same as another school.' (AC interview 1.14)

This rigorous approach to moderation also extends to the handover of assessment records between successive class teachers, thus supporting continuity of expectation between one year-group and the next:

> … at the end of the year, we have a changeover, where you'll sit with the teachers for the next year and you'll share through books and sort of agree, disagree, moderate together, so that you have got an agreement across all of it…' (AC interview 1.14)

Whilst the approach taken to evidencing and moderating teacher's summative judgements to monitor progress at school C appear to have maximised reliability, there is some evidence that this may have come at the expense of manageability, since the cross-referencing between tracking grids and samples of work was incomplete when monitored in June 2014.

School D: strong validity in progress tracking through pupil self-assessment

Pupil level

The SSL at School D identified a variety of opportunities provided for pupils to peer and self-assess work:

> Topic work is peer marked (P4), and the teacher would discuss with pupils what they should be looking for in the work that they assess (P2). At end of lessons, pupils self-assess both their understanding of procedures and knowledge (P3). This happens less when science is taught discretely, more so when science is linked to topic work. Self-assessment in KS1 is more informal but does happen. (SSL interview, 11.13)

At the end of lessons, pupils self-assess their work by commenting on their understanding of both procedures and concepts. Older pupils record this information in writing, younger pupils vocalise their thoughts for the teacher to record for them (P3). The self-assessment information gathered can be used by teachers to inform their judgements of pupil progress (M1).

Teacher level

The SSL explained that she wanted to see observation playing a more prominent role in teacher assessment (T3), particularly with older pupils. Pupils are asked to review and comment on formative feedback comments made by their teacher (T6). Feedback to KS1 pupils is given immediately, whereas with older pupils time is given for pupils to respond to comments made on their work during science lessons (T7). An example of a typical comment written by the teacher on an older pupil's work about the concept of water resistance was

> [You are] beginning to think why they [plasticene shapes] will be quicker or slower. Can you extend this further?

The sort of verbal feedback typically provided by a teacher of 5-6 year-old pupils in a lesson on properties of materials involved the teacher sharing with the class significant comments made by one pupil about what makes a material suitable for a particular purpose.

Monitoring level

Teachers at School D make judgements of pupils' attainment in science using a range of evidence collected in each child's topic-work book over the course of a unit of work in science (M1):

> Topic books gather a rich range of information for teachers to use in making judgements. Self-assessment is part of the process and is recorded in the topic books – in KS2 pupils score themselves out of five for various criteria and comment on their score. (AC interview, 1.14)

For younger pupils teachers make and record comments and observations (T3) principally on 'Post-it' notes. For older pupils some note-taking of comments also occurs although most assessment judgements are made through marking of written work (T4).

The close relationship between formative and summative assessment in School D was explained by the SSL:

> Assessments are made after each lesson, with summative judgements made at the end of a unit of work using these day-to-day assessments. Judgements are not just based around an end of unit task but, instead, are informed by the formative assessment information gathered. (SSL interview 11.13)

The SSL acknowledged that not all procedural understanding (science inquiry skills) can be recorded meaningfully or easily in written form in topic books, thus potentially limiting the validity of the summative judgements being made. However, the prominent role played by pupil self-assessment in progress monitoring increases the participant-informed validity of this process. During 2013-14 School D did not use consensus moderation to check or improve the reliability of teacher summative judgements.

School E: Enhancing reliability of teacher assessment through modeling performance and manageable moderation

Pupil level

Pupil focus on science objectives appears to be strong (P2) within weekly lessons, dedicated group or class books, classroom displays and regular science events. There is also a strong emphasis on speaking, listening and group work which is evident throughout the school: all observed lessons contained opportunities for pupils to explore their ideas with peers and staff. The emphasis on talk, combined with a focus on self and peer assessment (P3, P4), points to the way pupils are involved in the monitoring of their own learning. For example, in a Year 5

lesson on Space (13.1.14) the pupils were physically modeling the orbit of the Earth around the sun using different sized balls. As they moved the Earth ball they gave a commentary on what was happening, which was then peer-assessed for clarity and accuracy (P4). The groups gave advice to each other for how to improve their explanation. The teacher emphasised that to 'become 5* scientists' they should aim to use scientific vocabulary accurately, so the pupils listened out for the word 'orbit' in the explanations and watched to check that the Earth ball was moving whilst the sun remained still. Such use of peer assessment appears manageable because it all takes place within the lesson, and the focus on accurate use of scientific vocabulary provides clear criteria for reliable teacher and pupil judgements. However, by making the success criteria manageable for the pupils (e.g. to listen for the use of 'orbit') there is a risk that validity is compromised since it could be rote use of vocabulary rather than understanding which is being assessed. The teacher did address this issue at the end of the lesson by asking pupils to record their explanations in drawing and writing, providing a further assessment opportunity.

Teacher level

Elicitation of pupil ideas is appropriate to the age group (T1); for example, in Year 1 'floorbooks' are used to record pupil responses (a floorbook is a written record of pupils's utterances in the form of a large-format, 'home-made' book). A teacher or assistant sits with a group and scribes the pupils's comments in response to a stimulus, such as materials and magnets (T3). Such interactions could be time consuming to undertake with all groups - raising questions of manageability - but this depends on the classroom set up and availability of adults. The floorbooks contain an accurate record of the pupil utterances (T4) and the adult working with the group has been briefed on which key questions to ask to stimulate the discussion, thus enhancing consistency (reliability) of data collection. The school science scheme of work also contains success criteria (T2) and links to a progression of skills wheel. It could be questioned whether such clear expectations (supporting reliability) may reduce validity since the tight focus may lead to other aspects being ignored, or perhaps a 'tick box' culture where the teacher is waiting for a particular word rather than probing understanding across the topic.

8-9 year old pupils had found it difficult to construct branching identification keys for animals, so the teacher adapted the next lesson (observed 28.3.14) (T5) to begin with pupils making keys with 'Post-it' notes in small groups. After talk partners had formulated yes/no questions to divide the animals, pupils identified for themselves whether they felt confident in this activity or not (P3); data which the teacher used to group them. Pupils were given time to look at others' work (T7), being asked to pick out elements of a successful key before returning to improve their own key (P5). This was a manageable way for the pupils to give and act on feedback, with the teacher noting on her plan pupils who 'stood out' – those who struggled and those who exceeded expectations (T3). The class had constructed the success criteria for what constituted an effective branching key within the lesson, which supported reliability of pupil and teacher judgements.

Monitoring and summative reporting levels

Assessment at School E is based on a shared understanding of 'what good science looks like' (SSL interview 11.13) (M3). Each class displays science 'star' guidance and a 'skills wheel'. The skills wheel both supports coverage (M1) and tracks whole class skills development (S1) with wedges coloured in when the majority of the class fell confident. The science 'stars' provide details of, for example, what is expected of a 4* scientist so pupils know the criteria by which their work is judged (M4). Teachers use evidence gathered in pupil books or group floor-books - together with their observations of how the pupils have responded in lessons - to make a summative judgement at the end of each topic (M1, S3). Teachers' confidence in

matching these judgements to external criteria is supported by a series of brief (10-minute) science moderation sessions (M2), which take place within staff meetings across the year:

> Moderating regularly in small manageable chunks helps us to maintain a high profile for science, gives teachers confidence and means we have super evidence of pupils' attainment. (SSL interview 11.13)

These moderation meetings have contributed a portfolio of exemplar assessment evidence which supports a shared understanding of progression (M3) and reliability of judgement.

DISCUSSION AND CONCLUSIONS

Overall, our findings suggest that, in England, there is a wide variety of practice in primary science teacher assessment. This diversity has been encouraged by the UK government, which has resulted in a range of creative approaches to formative assessment of pupils' scientific skills and knowledge, from 'learning walls' to effective teacher feedback and peer assessment. However there is very little evidence of formative assessment being used to inform summative judgements (Nuffield Foundation 2012). Some schools remain wedded to bureaucratic numerical tracking systems with little evidence to support the validity or reliability of the judgements upon which pupil attainment levels are based.

The five snapshots of school practice in science assessment provided by the case studies above, whilst all differing in the tools used and the ways in which pupils' progress is tracked, display some common features which our analytical framework would suggest exemplify effective practice, including a concern to involve pupils as much as possible in assessing their own science progress, providing feedback to each other and responding to feedback from their teachers. This raises the issue of the use of group activities to make assessment judgements on individual pupils. The influence of others in the group – including the opportunities to verbalise thought and receive feedback from peers – is a fundamental principle of formative assessment, whereas summative assessment usually requires the evaluation of individual performance. Some schools are using a rigorous approach to evidencing and moderating teacher judgements; although this raises manageability issues it is nevertheless informing our developing synthesis.

Indeed, we might argue that the synthesis of the analytical framework (appendix 1) itself represents an answer to RQ3, in that its set of criteria together constitute a 'best practice' model of whole-school science teacher assessment. Whilst none of the case-study schools individually exemplifies all of the cells in the pyramid to an acceptable level of quality, between them they provide a picture of what a 'perfect' whole-school system would look like. The use of this framework as a self-evaluation tool has invited schools to measure themselves against this model, which many have found helpful. By exemplifying all the cells within the pyramid in an online version of the tool (see www.pstt.org.uk/SiteDocuments/doc/taps/taps-pyramid-final.pdf) we have synthesised the most effective aspects of teacher assessment in science from all participating schools in the TAPS project.

REFERENCES

Black, P., Harrison, C. Lee, B., Marshall, and Wiliam, D. (2003). *Assessment for Learning: putting it into practice*. Maidenhead: OUP.

Black, P., Harrison, C, Hodgen, J. Marshall, B. and Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy and Practice 18* (4), 451–469.

Brill, F. and Twist, L. (2013). *Where Have All the Levels Gone? The Importance of a Shared Understanding of Assessment at a Time of Major Policy Change*. Slough: NFER.

Daugherty, R. (2009). National Curriculum Assessment in Wales: Adaptations and Divergence. *Educational Research*, *51*(2), 247-250.

Davies, D., Collier, C. and Howe, A. (2012). Assessing Scientific and Technological Enquiry Skills at Age 11 using the e-scape System. *International Journal of Technology and Design Education, 22*, 247-263.

Earle, S. (2014). Formative and summative assessment of science in English primary schools: evidence from the Primary Science Quality Mark. *Research in Science & Technological Education*, *32(*2), 216-228.

Edwards, F. (2013). Quality Assessment by Science Teachers: Five Focus Areas. *Science Education International, 24* (2), 212–226.

Gipps, C., and Murphy, P. (1994). *A Fair Test? Assessment, Achievement and Equity.* Buckingham: OUP.

Harlen, W. (2013). *Assessment & Inquiry-Based Science Education: Issues in Policy and Practice*. Trieste: Global Network of Science Academies (IAP) Science Education Programme (SEP).

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education, 28*(1), 91-105.

Kelly, L., and Stead, D. (eds.) (2013). *Enhancing Primary Science.* Maidenhead: OUP, McGraw Hill.

Klenowski, V. (2011). Assessment for learning in the accountability era: Queensland, Australia. *Studies in Educational Evaluation 37*(1), 78-83.

Klenowski, V. and Wyatt-Smith, C. M. (2010). Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. *Assessment Matters, 2*, 107–131.

Mansell, W. and James, M. (2009). *Assessment in schools: fit for purpose?* London: Teaching and Learning Research Programme.

Newton, P. (2009). The Reliability of Results from National Curriculum Testing in England. *Educational Research, 51* (2), 181–212.

Nuffield Foundation (2012). *Developing Policy, Principles and Practice in Primary School Science Assessment.* London: Nuffield Foundation.

Sahlberg, P. (2011). Lessons from Finland. *Education Digest, 77*(3), 18-24.

Stanley, G., MacCann, R., Gardner, J., Reynolds, L. and Wild, I. (2009). *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development.* Oxford: Oxford University Centre for Educational Assessment.

Stobart, G. (2012). Validity in Formative Assessment. In J. Gardner (ed.) *Assessment and Learning: Second Edition.* London: Sage.

Wiliam, D. (2003). National Curriculum Assessment: How to Make It Better. *Research Papers in Education, 18* (2), 129–136.

Wiliam, D. and Black, P. (1996). Meaning and Consequences: A Basis for Distinguishing Formative and Summative Functions of Assessment? *British Educational Research Journal, 22* (5), 537-52.