Sarah Earle (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

[Accepted version]

Abstract

Background: Since the discontinuation of Standard Attainment Tests (SATs) in science at age 11 in England, pupil performance data in science reported to the UK government by each primary school, has relied largely on teacher assessment undertaken in the classroom.
Purpose: The process by which teachers are making these judgements has been unclear, so this study made use of the extensive Primary Science Quality Mark (PSQM) database to obtain a 'snapshot' (as of March 2013) of the approaches taken by 91 English primary schools to the formative and summative assessment of pupils' learning in science.
Programme description: PSQM is an award scheme for UK primary schools. It requires the science subject leader (co-ordinator) in each school to reflect upon and develop practice over the course of one year, then upload a set of reflections and supporting evidence to the database to support their application. One of the criteria requires the subject leader to explain how science is assessed within the school.
Sample: The dataset consists of the electronic text in the assessment section of all 91 PSQM primary schools which worked towards the Quality Mark in the year April 2012 to March 2013.
Design and methods: Content analysis of a pre-existing qualitative dataset. Text in the assessment section of each submission was first coded as describing formative or summative processes, then sub-coded into different strategies used.
Results: A wide range of formative and summative approaches were reported, which tended to be described separately, with few links between them. Talk-based strategies are widely used for formative assessment, with some evidence of feedback to pupils. Whilst the use of tests or tracking grids for summative assessment is widespread, few schools rely on one system alone. Enquiry skills and conceptual knowledge were often assessed separately.
Conclusions: There is little consistency in the approaches being used by teachers to assess science in English primary schools. Nevertheless, there is great potential for collecting evidence that can be used for both formative and summative purposes.

Keywords: Assessment, primary, science, formative, summative

Introduction

The curriculum as experienced by children is shaped by assessment practices; thus it is essential for such practices to be well understood by teachers. Currently, primary teachers in England are required by law to allocate an assessment level in science to each child at ages 7 and 11. Since the removal of Standard Attainment Tests (SATs) in 2009, these level judgements have relied upon teacher assessment. Whilst many teachers do not regret the removal of SATs, the subsequent increased emphasis on making reliable teacher assessment judgements has caused concern (Turner et al 2013: 3). Gardner et al (2010) argue that teacher assessment is a more valid means of summative assessment than testing because it can be based on the wider range of evidence available to teachers in the classroom, for example, observations, discussions and lines of enquiry. Teacher judgement can take into account a range of outcomes which are not easily assessed in a test; this is particularly important for science since its essence is practical, scientific enquiries can utilise dialogue, collaboration, practical skills and problem solving in real life contexts (Kelly and Stead 2013). Nevertheless, whilst validity may be stronger than for tests, questions remain regarding the reliability of teacher assessment (Harlen 2007:25, Black et al 2011), since teachers can find such summative judgements difficult to make, and also because there are limited opportunities for comparing their judgements with others'. However, Wiliam (2003) argues that teacher assessment can be made more reliable, and that there is inevitably a 'trade off' between reliability and validity. With large-scale collection of evidence and effective moderation procedures, where teachers compare and discuss judgements, reliability of summative teacher assessment can be as high as it needs to be (Harlen 2007), though this raises issues of manageability. Overall, a major concern raised by the current situation is the lack of centralised guidance for primary teachers on how to assess science. If teachers do not have an explicit view of what makes 'good' assessment in science, then it becomes difficult to decide how to make improvements in practice (Gardner et al 2010:8), there may be poor 'teacher assessment literacy' (Edwards 2013). With "no single approach to teacher assessment" (Harlen 2012: 137) and researchers noting the 'formidable challenge' (Black 2012: 131) of developing classroom assessment practices, there is a distinct lack of clarity in this area, which has opened the door to a plethora of home-grown and commercially-produced 'solutions'.

This lack of clarity led the author to undertake a content analysis of an existing dataset in order to take a 'snapshot' of current approaches to teacher assessment of science being used by a sample of 91 primary schools in England. This could then be used to identify common strategies with their associated strengths and weaknesses and form the basis for disseminating effective assessment practice more widely. The study made use of written submissions made by school science subject leaders to the Primary Science Quality Mark (PSQM) database. All participating schools have been informed that submissions may be used anonymously for research purposes. The Primary Science Quality Mark is an award scheme to enable primary schools '*to evaluate, strengthen and celebrate their science provision'* (psqm.org.uk). It requires the science subject leader (co-ordinator) in each school to reflect upon and develop practice over the course of one year, then upload a set of reflections and supporting evidence to the database to support their application. The Quality Mark is awarded at Bronze, Silver or Gold, after consideration of 13 criteria including subject management, teaching, learning and assessment approaches. One of the 13 criteria (C2) requires the subject leader to explain how science is assessed within the school, so it was analysis of the evidence submitted under criterion C2 that formed the basis of this study. A particular focus of the analysis was how teachers described their approaches to formative and summative assessment in science, since a closer relationship between these is seen by some as crucial to the effective deployment of teacher assessment in tracking pupil progress (Wiliam and Black 1996, Hodgson and Pyle 2010, Nuffield Foundation 2012, Harlen 2013).

The relationship between formative and summative assessment in primary science education

The distinctions between formative and summative purposes of assessment have received much attention in the UK during the last 15 years, with the importance of formative assessment stressed by renaming it 'Assessment *for* Learning (AfL)' (Black and Wiliam 1998), an: "ongoing planned process that focuses on identifying the next steps for improvement" (Harrison and Howard 2009: 28). AfL requires the active involvement of children and researchers stress the importance of dialogue and questioning (Black and Harrison 2004). By contrast, summative assessment has been termed 'Assessment *of* Learning (AoL) ' (Black and Wiliam 1998), since it aims to summarise pupils' learning for the purpose of accountability, taking a "snapshot in time of their performance" (Mawby and Dunne 2012: 139). Such summaries of learning - either grades or narratives - can be reported for example, to parents, other teachers, school leadership teams or school inspectors. In recent years mounting evidence for the positive impact of formative assessment on children's learning (Hattie 2006, Gardner et al 2010) has elevated the status of AfL, whilst evidence demonstrating the harmful effects of high stakes summative testing (Newton 2009) and its distorting effects on the taught curriculum (Wiliam 2003) has led some teachers to view AfL and AoL as the 'good' and 'bad' sides of assessment respectively (Harlen 2013) .

However, in practice it is sometimes difficult to draw clear distinctions between AfL and AoL (Davies et al 2012), since the same assessment tasks may be used for both summative and formative purposes (Hodgson and Pyle 2010), e.g. the formative use of summative tests (Black 2003). Harlen (2007) states AfL and AoL differ only in purpose and degree of formality, which suggests that rather than a dichotomy, it may be more useful to see these assessment processes as dimensions (Harlen 2013) or perhaps a continuum (Wiliam and Black 1996). Harlen (2013) asserts that any assessment opportunity can be used for formative or summative purposes, thus it is the purpose rather than the strategy which decides the label. Advocates of change in assessment practices suggest that it is possible and desirable to use the same evidence for both formative and summative purposes (Nuffield Foundation2012). 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. This does not mean doing formative and summative assessment at the same time, for example, when marking work it is not helpful to put a summative score as well as comments for improvement, since these comments are likely to be ignored if there is also a score (Wiliam 2011). However, if the evidence compiled from everyday interactions in the classrooms can be aggregated into a summary statement or level then the negative impact of summative testing could be avoided. There is not universal agreement that this is the way forward in assessment since there are those who argue that: "any attempt to use formative assessment for summative purposes will impair its formative role" (Gipps 1994:14). Wiliam and Black (1996) argue that this is possible as long as the elicitation of evidence is separated from the interpretation or judgement. Harlen (2007) also asserts that *"it is essential to* *ensure that it is the evidence used in formative assessment and not the judgements that are summarised"* (p. 117).

Will a blurring of the lines between formative and summative assessment support practitioner understanding? Brill and Twist (2013) highlight the importance of teachers developing a shared, secure understanding of assessment, particularly in a time of change in assessment policy. There is evidence that some teachers in the UK are misinterpreting AfL to mean frequent testing, demonstrating a lack of understanding of the aims of assessment practices (Black 2012). Swaffield (2011: 433) also questions whether AfL and formative assessment are synonymous, questioning the: "distorted practices that are erroneously termed AfL" in government policy (DCSF 2008). This study aims to consider which assessment practices are used for primary science and the degree of separation of formative and summative assessment in practice.

Method

This research employs content analysis of a pre-existing dataset: the submissions to an online database of science subject leaders in all 91 English primary schools who worked towards the PSQM in Round 4 (April 2012 to March 2013). Each PSQM Round begins in either September or April and lasts for one year, while the subject leader receives training, audits school practice, develops and implements an action plan, finally gathering evidence and reflecting on the impact of actions. Round 4 evidence was the most recent available at the time of analysis so provided the most up-to-date 'snapshot' of practice. Data consisted of written reflections in Spring 2013 regarding current school practice in science and developments over the past year. The C2 reflections from all 91 schools have been used to catalogue the types of formative and summative assessment currently being used. It is important to note that this sample of schools have put themselves forward for an award and thus may be developing practice at a different rate to other primary schools in England. Bronze schools would be using the award as a way of receiving training and raising the profile of the science in the school, Silver schools aim to develop good practice across the school and Gold schools would aim to share good practice beyond the school. Therefore PSQM schools would perhaps be more likely to be evaluating and developing their assessment practices. At this time the teachers knew there would be a new curriculum for September 2014 and may have seen the draft in early form but at the point of submission there had been no details about new assessment guidance.

The subject leader reflections consisted largely of descriptions of the assessment strategies which were being trialled or used across the school. Analysis of such summaries for this study led to consideration of the proportion of schools using different strategies, since it was recognised that judgement of 'teacher assessment literacy' (Edwards 2013), would require a richer dataset; this is one of the aims of the next stage of the research within the Teacher Assessment in Primary Science (TAPS) project funded by the Primary Science Teaching Trust. In order to build a numerical picture of the types of assessment being used by the 91 schools the C2 reflections were coded using a qualitative analysis software called Atlas.TI which supports the creation and organisation of coded extracts. Simple key word frequencies were not suitable, since subject leaders discussed the merits of different strategies, thus it is important to consider the coding decisions in a little more detail below. To separate formative and summative methods of assessment, it was important to clearly identify a practical definition of ‘summative’ which could be applied consistently to this data set. The method was classified as summative if:

- it was described as ‘end of unit’ or ‘end of year’
- it fulfilled a summarising purpose e.g. passed onto the next teacher or put into the school tracking software(where a level or sublevel judgement may be assigned to each child to enable staff to track numerical progress since the last data entry point)
- it was identified by the teacher as ‘summative’

Formative assessment was harder to classify, partly due to the wide range of methods being employed. There is also the question of whether the strategies described were being used as AfL to identify the next steps for the learner. AfL is: “not simply a matter of teachers adopting assessment for learning strategies” (Harrison and Howard 2009: 32); the information gained should lead to an impact on learning by adaption of learning experiences. For the purposes of comparing methods – whether or not they were explicitly identified as supporting learning - they were termed ‘elicitation strategies’ (Harlen and Osborne 1985, Ollerenshaw and Ritchie 1997). The wide range of elicitation strategies described across the 91 schools led to consideration of how to categorise them. Following Wiliam and Black (1996), the analysis attempted to separate the collection of assessment evidence from teacher judgement, an important consideration if exploring the possibility of using the information gathered for both formative and summative purposes. Some elicitation strategies were classified as primarily judgemental, such as teacher marking or annotating work, and self or peer evaluation. Observation and questioning were harder to classify, it could be argued that they both involve collecting rather than judging evidence. But in recording the observation (e.g. by note taking on post its or photographing) or deciding what question to ask next, the teacher is inevitably making a selection, which involves a judgement about the child’s learning and, in the case of questioning, potentially intervening. Since the mention of these techniques in a science subject leader’s summary is insufficient to separate the two purposes, they have both been included in the elicitation data for completeness.

Findings

Summative Assessment

The categorisation of summative assessment methods can be seen in summary form in Figure 1 and in more detail in Figure 2. Analysis of statements from the 91 subject leaders found that only two did not explain how they assessed science summatively, thus the percentages in this section are based on 89 schools. Many schools (38%) mentioned testing, but only 10% of these used testing alone (see Fig 1). The others used test results as part of the information, combining this information with other methods such as tracking grids.

One form of tracking grid mentioned by 36% of schools was Assessing Pupil Progress (APP), introduced by the UK Department for Children, Schools and Families (DCSF 2010), but no longer government policy. These grids provide detailed assessment criteria which can be highlighted when a child or group is deemed to have met a particular criterion. A range of associated benefits of using the APP approach were mentioned by several subject leaders:

> *Science APP not only allows the head teacher, staff and myself to track pupils’ progress but it has also helped to maintain the high profile of science in our school following its removal from SATs. It*

> *also informs planning and is a valuable tool for ensuring effective differentiation in the classroom.* (extract from subject leader submission)

> *The impact of introducing Science APP has been that staff feel more confident assessing science, assessment is consistent across school, and gives a good overview of a child's learning and progress in science rather than relying on a snapshot 'test-style' assessment.* (extract from subject leader submission)

Several schools had adapted the APP grids, for example, by rephrasing criteria in the form of 'I can…' statements for pupil self-assessment at the end of units or developed their own tracking grids containing levelled criteria. As with testing, whilst 36% of schools were using APP tracking grids, and a further 20% using other tracking grids (commercial or of their own construction) only around a third of these were using APP alone. The proportion using 'other' tracking grids *alone* was much higher (85%), possibly because these included conceptual as well as procedural knowledge, whilst APP is exclusively skills-focused. Since at this time teachers were required to report attainment levels for both scientific knowledge and skills it appears that there was a tendency to use separate systems for these components: typically testing for knowledge and APP for skills:

> *APP is used by all staff to assess pupil's Sc1 understanding and skills. In addition to this, colleagues use Mini Sats to assess pupils' knowledge and understanding in Science* (extract from subject leader submission).

One surprising feature of the data regarding APP was that, although several submissions expressed concern over its manageability as a strategy for tracking pupil progress in science – added to which it only covers enquiry skills, is no longer government policy and is not compatible with the changes to the national curriculum in 2014 – some submissions were still considering its introduction, as in the following example:

> *Our school has been using Maths and English APP for several years. APP for Science has not been introduced. I have discussed it briefly with our Headteacher but at the time it was considered too much added pressure for staff… I am considering trialling using APP in the summer term [when pressure of SATS is gone!] I am aware that this is a major area for development personally and school wide.*

Formative Assessment

As discussed above, the assessment techniques analysed at this stage will be termed 'elicitation' strategies rather than formative strategies, since whilst we can assume they have been used to find out what the children know or understand, there is often not enough explanation to judge if they fulfil a formative purpose; explicit formative use will be discussed in the next section. Data indicated a wide range of elicitation strategies being used in the 91 schools, from paper-based tests to pupils raising their own questions. Figure 3 groups together similar approaches to elicitation in science, such as teacher-led talk, collaborative activities, observation and paper/task-based, such as KWL grids in which children record what they Know, Would like to know and, at the end of the unit, what they have Learnt. These elicitation strategies range in terms of how open or closed the tasks are. For example, a mind map where the child records what they know about forces was classified as an open task whilst a true/false quiz was deemed closed. Other variables were difficult to categorise from the subject leaders' reflections, for example whether the elicitation was pupil-led or teacher-led, or whether the children were working individually or collaborating on some tasks. For example, whilst role-play tends to involve collaboration and presentations were mentioned by five schools, it was not clear whether the children were working alone or in a group. Whilst eight schools mentioned the use of concept cartoons (Naylor and Keogh 200), they did not say whether these are used to stimulate a class discussion or for individual responses. Talk featured strongly as an elicitation strategy; for example seven schools mentioned the use of pupil talk partners to discuss ideas in pairs. However, the use of 'questioning' by 29 schools was unclear , since this could have involved individuals, groups or the whole class; in the form of fast-paced closed questioning or open-ended consideration of 'big' questions such as 'what would life be like without friction?' Nevertheless, despite the ambiguous nature of some of the terms, it is clear that schools were collecting a wide range of evidence of pupils' science learning, both long-lasting and ephemeral (Wiliam and Black 1996).

There is evidence that some schools involve pupils to monitor their own learning in science. 36% of schools mentioned self-assessment and 8% peer-assessment. A closer look at the descriptions of self assessment (Figure 4) reveal that whilst eight reported only that pupils were 'given the opportunity' to self assess, those who were more specific fell into three groups. 10 of the schools reported asking pupils to assess their own performance against stated learning objectives. These pupils were evaluating their work by drawing 'smiley faces' if they felt they had met objectives; colouring 'traffic lights' red, amber or green or putting their thumbs up, sideways or downwards to indicate their level of understanding; ticking the learning objective or the success criteria in their written work; or identifying their next steps or 'wish' for their science learning. Nine schools were asking pupils to consider their progress by highlighting 'I can' statements, learning ladders, APP grids or level checklists.

28 schools identified feedback from teachers to pupils by marking or annotating work, although it is likely that this is an underestimation since marking is such a day-to-day routine for teachers that respondents may not have seen it as a separate assessment strategy. Exactly how 'marking' was described merits further analysis since if subject leaders noted pupils acting on the teacher's written advice it would suggest that they are being formative, with assessment being used to support learning, however the formative drive would be reduced if work was being annotated to provide evidence for accountability. Of the 25 schools specifically mentioning 'marking', nine emphasised teacher judgement - for example, highlighting the learning objective to show that it has been achieved - whilst the other 16 went on to describe how they use marking to move pupils' learning forward by explaining their next step, asking challenging questions or identifying 'two stars and a wish' where two features are celebrated and one provided as a next step. Such 'feed-forward' marking suggests that AfL is taking place, provided that children are given time to respond to the marking comments (Harrison and Howard 2009). A further 10 schools described using elicitation evidence to identify gaps in learning and then alter their planning or provide additional tasks for the children. An additional five schools, bringing the total identifying AfL strategies to 31, described how they move pupils' learning forward by prescribing 'next steps', for example on a 'working wall' on which pupils could compare their work to success criteria or level checklists. Black et al (2003: 78)

would perhaps question the use of levels here, suggesting that pupils who are given feedback as marks negatively compare themselves with others (ego-involvement) and ignore comments, whilst comment-only marking helps them to improve (task-involvement). It is however possible that these schools are using the level descriptors as a way of supporting children to know what good quality work 'looks like' (Black and Harrison 2004: 4).

Discussion

The separation of scientific skills and knowledge, particularly in relation to summative assessment, is a strong feature of the data reviewed above which supports other research findings (e.g. Hodgson and Pyle 2010). 37% of schools in this sample described a separation of assessment methods, for example, using tests for conceptual understanding and tracking grids for procedural understanding. Although there is agreement in the literature that both conceptual and procedural knowledge should be assessed (Howe et al 2009), the majority of assessment research is concerned with developing science concepts rather than skills (Hodgson and Pyle 2010, Black and Harrison 2004) and when skills have been addressed they are considered separately from concepts (e.g. Russell and Harlen 1990). The importance of pupil talk and effective questioning to support AfL has been well documented (e.g. Alexander 2006), but again it is the development of science concepts which dominate (Earle and Serret 2012). The use of separate systems raises questions of manageability for teachers, especially once the extensive requirements for assessment of English and Mathematics are taken into account. It also raises more fundamental questions about how primary school assessment is representing the nature of science and whether it is possible or desirable to separate knowledge and skills in this way. The revised national curriculum in England advises that: "working scientifically… must always be taught through, and clearly related to, substantive science content in the programme of study" (DfE 2013a: 5). Nevertheless, those who favour tick-list style tracking documents such as APP would argue that it is necessary to identify specific scientific skills from an activity which may also have conceptual content, for example, noting whether a child observes closely when exploring the translucency of a fabric with a torch.

The reported use of APP provides an interesting comparison with an earlier summary of Round 1 PSQM data collected in 2011 (Turner et al 2013), in which from a sample of 37 schools, 25 of them (68%) were using APP. This analysis of Round 4 data suggests a dramatic drop in the use of APP over a two-year period, with only 13% solely reliant on this approach to tracking achievement, although a further 24% were using it in combination with other methods, as discussed above. Political context is an important factor here: Round 1 schools were working towards the Quality Mark between April 2010 and March 2011, only one year after the removal of SATs testing: *'The reflections on assessment submitted by the majority of subject leaders focused on the problem of filling the gap left by removing the science SAT'* (Turner et al 2013: 22-23). APP had been disseminated via the National Strategies in the Summer of 2010 and, although non-statutory, many of the Round 1 schools were in the process of trying it out. By the time of the Round 4 submissions the new government had 'archived' the APP supporting materials on their website: *'APP will continue as a voluntary approach to pupil tracking and whilst many schools may find it useful, it is for the school to decide if they want to use it or not. There are no plans to make APP statutory or to introduce it for other subjects.'* (DfE 2011). Nevertheless, it is interesting to note that at least five schools in the sample were planning to introduce APP as a next step in their development of assessment procedures. Despite the government's ambivalent attitude towards APP, it appears some schools find it a useful tool, and others will try it out, despite their own worries, perhaps because of the lack of an alternative.

Subject leaders contributing to the Round 4 data devoted a considerable proportion of their reflections against criterion C2 to describing the development and monitoring of formative assessment strategies in science, suggesting that this had been a focus for development in many of the schools. Those who question whether schools are misinterpreting AfL to mean frequent testing (e.g. Black 2012 and Swaffield 2011) would be pleased to find that the schools in this sample did not appear to be over-using tests, or seeing testing as the only reliable form of assessment (Harrison and Howard 2009). They were using a wide range of strategies for eliciting children's ideas and at least one third appear to be using this information formatively to move the children's learning forward by, for example, adapting teaching or identifying next steps. Harrison and Howard (2009:1) assert that AfL, with its focus on promoting learning, has wide international currency, whilst summative assessment is more country-specific since this is more dependent on the particular framework for assessment. With popular UK primary science publishers such as Millgate House (e.g. Naylor and Keogh 2000) producing guidance for AfL, this may have helped subject leaders feel more confident in this area, whilst a general lack of guidance in summative assessment, apart from commercially-produced 'levelling tests' and the waning APP, had left teachers without a clear direction. To have separate systems for formative and summative assessment, and for the assessment of knowledge and enquiry skills, places an unmanageable burden on teachers (Harlen 2013).Thus many schools in the sample were keen to review their approach to science assessment, recognising that their current systems were not sustainable. With the advent of a new national curriculum with an assessment framework no longer level-based (DfE 2013b), this recognition of the need for change was well-placed (Nuffield Foundation 2012).

Conclusion

There is "no single approach to teacher assessment" (Harlen 2012: 137). Whilst some schools in the sample reported using APP or testing, a large number used more than one method for summative assessment and this was usually described separately from formative assessment strategies. Should we be worried about such a wide variety of practice? Perhaps not, as current UK government guidelines suggest that each school should choose its own assessment structures (DfE 2013b). Harrison and Howard (2009) suggest that: "it is consistency of principle not uniformity of practice that works". Thus variety is not a problem, as long as methods are based on a secure understanding of assessment purposes, identifying whether the aim is formative or summative. And is there secure understanding? The evidence here is inconclusive. Of course it is also important to remember that this sample is not representative of all English primary schools since they were working towards the Primary Science Quality Mark which required them to reflect upon, and perhaps develop, their assessment practices. So it is likely that other primary schools may have less developed assessment practices. The next stage of this research, within the TAPS project, will be to work with primary schools to develop a model for the assessment of science, exploring whether formative assessment can be used for summative purposes. The assessment model should support teachers' understanding of assessment, enabling them to collect valid and reliable data, using manageable processes, to support teaching and learning, and to facilitate formative and summative judgements.

Acknowledgements: The author wishes to thank the Primary Science Quality Mark for supporting access to their database of award submissions and the Primary Science Teaching Trust for funding the next stage of research, the Teacher Assessment in Primary Science (TAPS) project.

References

Alexander, R. (2006) Towards Dialogic Teaching – rethinking classroom talk. Cambridge: Dialogos

Black, P. and Wiliam, D. (1998*) Inside the black box.* London: GL Assessment

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

Black, P. and Harrison, C (2004) *Science inside the Black Box*. London: GL Assessment

Black, P. (2012) Formative assessment and learning. In Oversby, J. (Ed) *ASE Guide to Research in Science Education*. Hatfield: ASE

Briggs, M., Woodfield, A., Martin, C. and Swatton, P. (2008). *Assessment for learning and teaching in primary schools.* Exeter : Learning Matters

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* (NFER Thinks: What the Evidence Tells Us). Slough: NFER

Davies, D. and McMahon, K. (2003) Assessment for inquiry: supporting teaching and learning in primary science, Science Education International, 14, 4, 29-39

Davies, D., Collier, C., McMahon, K. and Howe, A. (2010) E-scape Assessment. Primary Science, 115, 18-21

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.

Department for Children and Families DCSF (2008) *The Assessment for Learning Strategy.* Nottingham: DCSF Publications

DfE (2011) http://www.education.gov.uk/schools/toolsandinitiatives/cuttingburdens/b0075738/reducing-bureaucracy/progress

Department for Education (2013a) *Science - Programmes of study for Key Stages 1-2.* London: DfE

DfE (2013b) http://www.education.gov.uk/schools/teachingandlearning/curriculum/nationalcurriculum2014/a00225864/assessing-without-levels (Accessed 14.6.13)

Earle, S. and Serret, N. (2012) Children communicating science. In Dunne, M. and Peacock, A. (Ed) Primary Science: A Guide to Teaching Practice, London: Sage

Edwards, Frances (2013) Quality assessment by science teachers: Five focus areas. *Science Education International,* 24, 2, 212-226

Gardner, J., Harlen, W., Hayward, L., Stobart, G. with Montgomery, M. (2010) Developing teacher assessment. Maidenhead: OUP

Gipps, C. and Murphy, P. (1994) A fair test? Assessment, achievement and equity. Buckingham: OUP

Harlen, W. (2000) The Teaching of Science in Primary Schools, (Third Edition), London: David Fulton

Harlen, W. (2006) On the relationship between assessment for formative and summative purposes. In Gardner, J. (Ed) Assessment and Learning, London: Sage

Harlen, W. (2007*) Assessment of learning*. London: Sage

Harlen, W. (2012) What research tells us about summative assessment. In Oversby, J. (Ed) *ASE Guide to Research in Science Education*. Hatfield: ASE

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

Harlen, W. and Osborne, R. (1985) A model for learning and teaching applied to primary science. *Journal of Curriculum Studies,* 2(17): 133-146.

Harrison, C. (2012) Assessment for learning: classroom practices that engage a formative approach. In Oversby, J. (Ed) *ASE Guide to Research in Science Education.* Hatfield: ASE

Harrison, C. and Howard, S. (2009)*Inside the Primary Black Box.* London: GL Assessment

Hodgson, C. and Pyle, K. (2010) *A literature review of Assessment for Learning in Science*. Slough: Nfer

Howe, A., Davies, D. McMahon, K., Towler, L., Collier, C. and Scott, T. (2009) *Science 5-11: A Guide for Teachers* (2nd edn.) London: David Fulton

Johnston, J. (2012) Deciding paradigms and methodology. In Oversby, J. (Ed) ASE Guide to Research in Science Education, Hatfield: ASE

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

Mawby, T. and Dunne, M. (2012) Planning for Assessment for Learning. In Dunne, M. and Peacock, A. (Ed) (2012) Primary Science: A Guide to Teaching Practice, London: Sage

Murphy, P. (Ed) (1999) Learners, Learning and Assessment. London: Sage

Naylor, S. and Keogh, B. (2000) *Concept Cartoons in Science Educatio.,* Sandbach: Millgate House

Naylor, S., Keogh, B. and Goldsworthy, A. (2004) *Active Assessment: Thinking, Learning and Assessment in Science*. London: David Fulton

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

Ollerenshaw, C. and Ritchie, R. (1997) (2nd ed.) *Primary Science: Making it Work.* London: David Fulton.

Russell, T. and Harlen, W. (1990) *Assessing Science in the Primary Classroom: Practical Tasks*. London: Paul Chapman Publishing

Swaffield, S. (2011) Getting to the heart of authentic Assessment for Learning, *Assessment in Education: Principles, Policy and Practice*, 18:4, 433-449

Webb, M. & Jones, J. (2009) Exploring tensions in developing assessment for learning, *Assessment in Education: Principles, Policy & Practice*, 16:2, 165-184

Wiliam, D. and Black, P. (1996) Meaning and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal* Dec96, Vol. 22 Issue 5, p537, 12p

Wiliam, D. (2003). 'National curriculum assessment: how to make it better', *Research Papers in Education*, 18, 2, 129–136.

Wiliam, D. (2011) *Embedded formative assessment*. Bloomington: Solution Tree Press

Turner, J., Keogh, B., Naylor, S. and Lawrence, L. (2011) *It's Not Fair, or is It?* Sandbach: Millgate House

Turner, J. with S. Marshall, A. Farley and L. Harriss (2013) *Primary Science Quality Mark: Learning from good practice in primary science.* London: Wellcome Trust

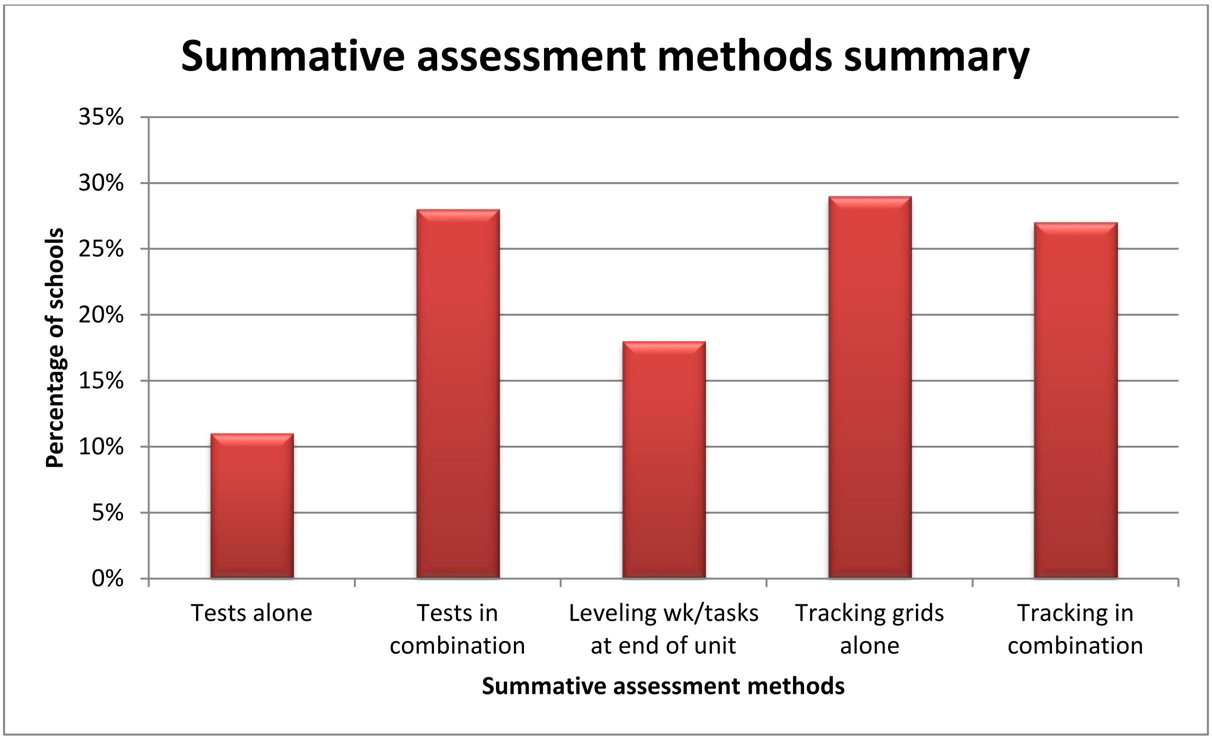

Figure 1: Summative assessment (summary) for PSQM Round 4 (March 2013, 89 schools since 2 did not specify)

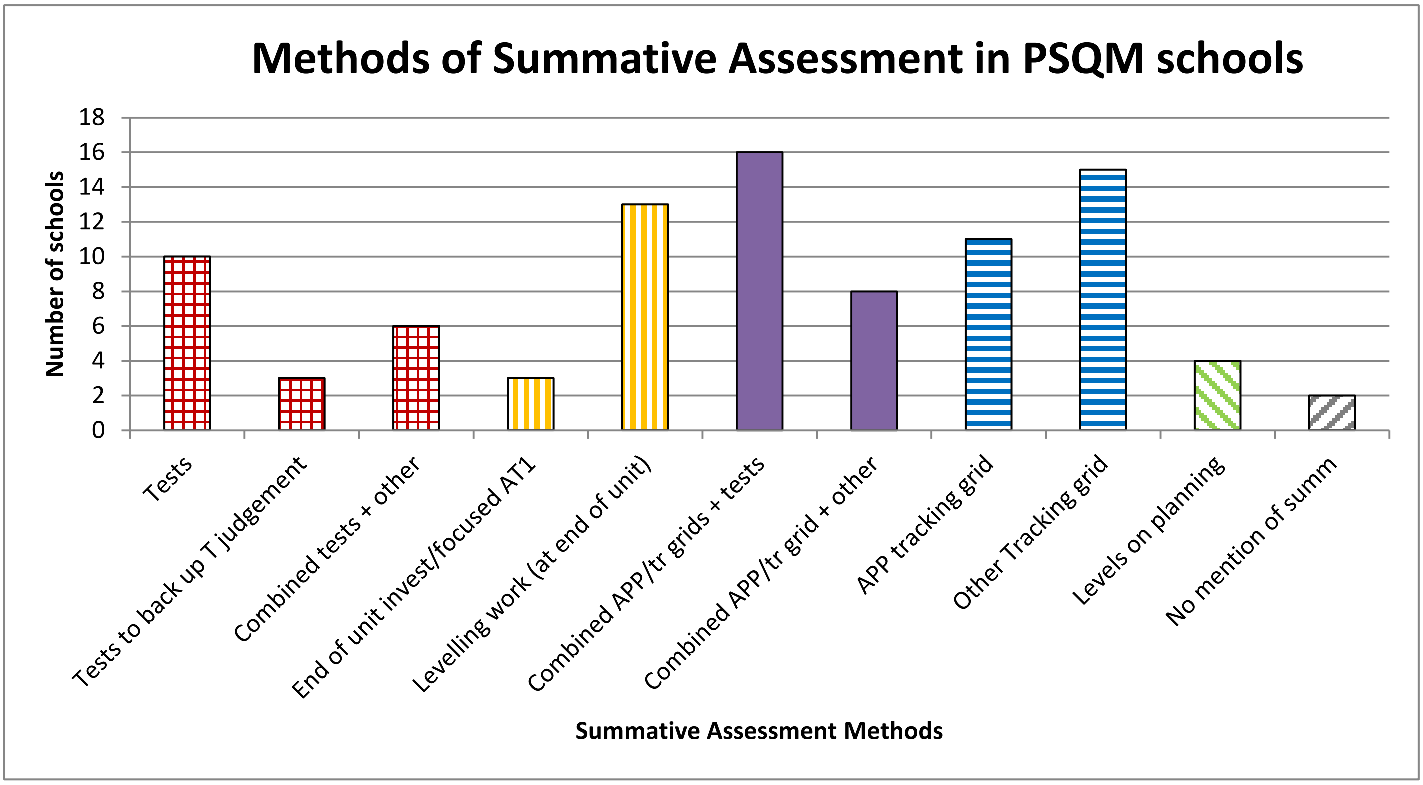

Figure 2: Summative assessment (detailed) methods for PSQM Round 4 (March 2013, 91 schools)

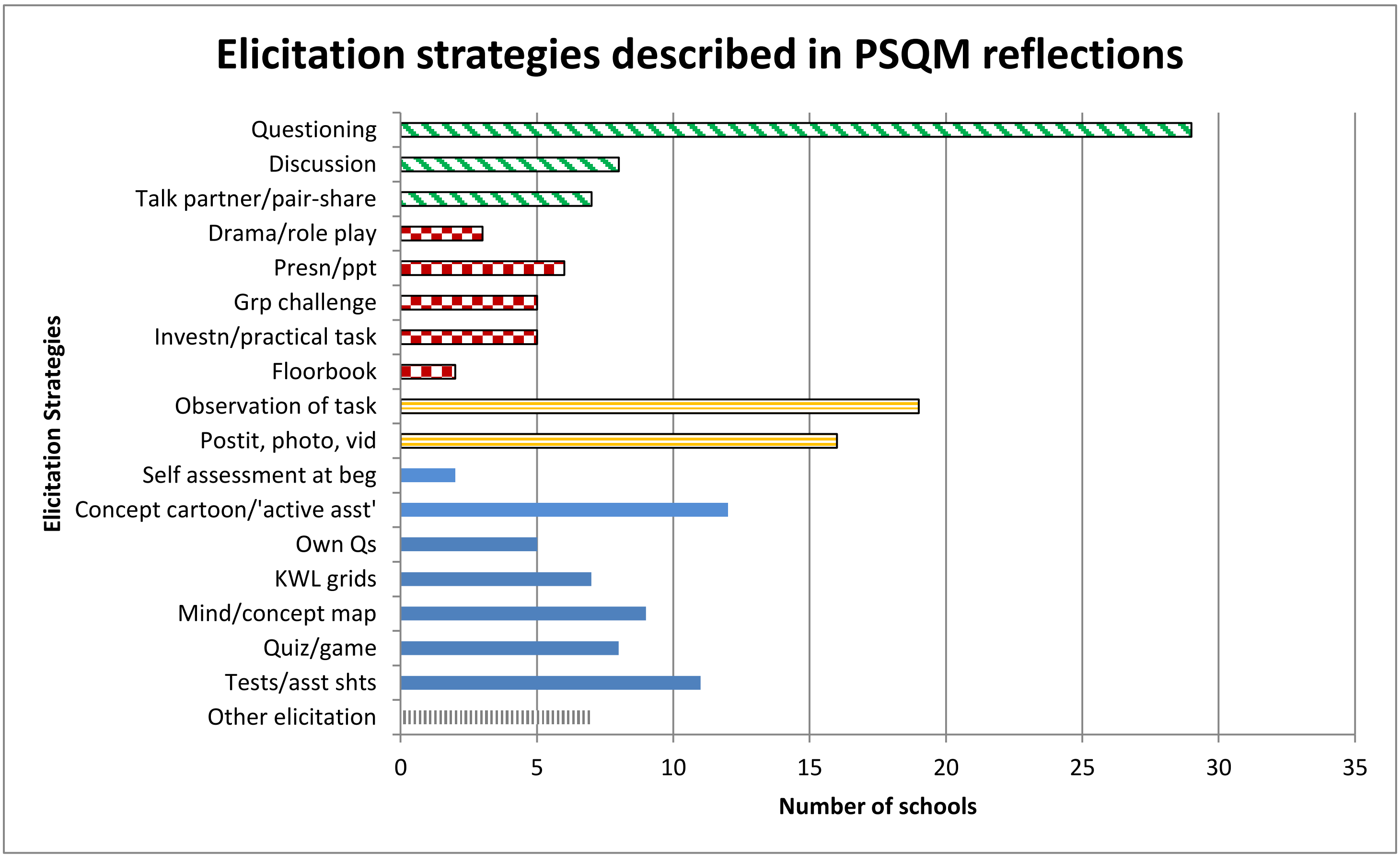

Figure 3: Elicitation strategies mentioned in reflections for PSQM Round 4 (March 2013, 91 schools)

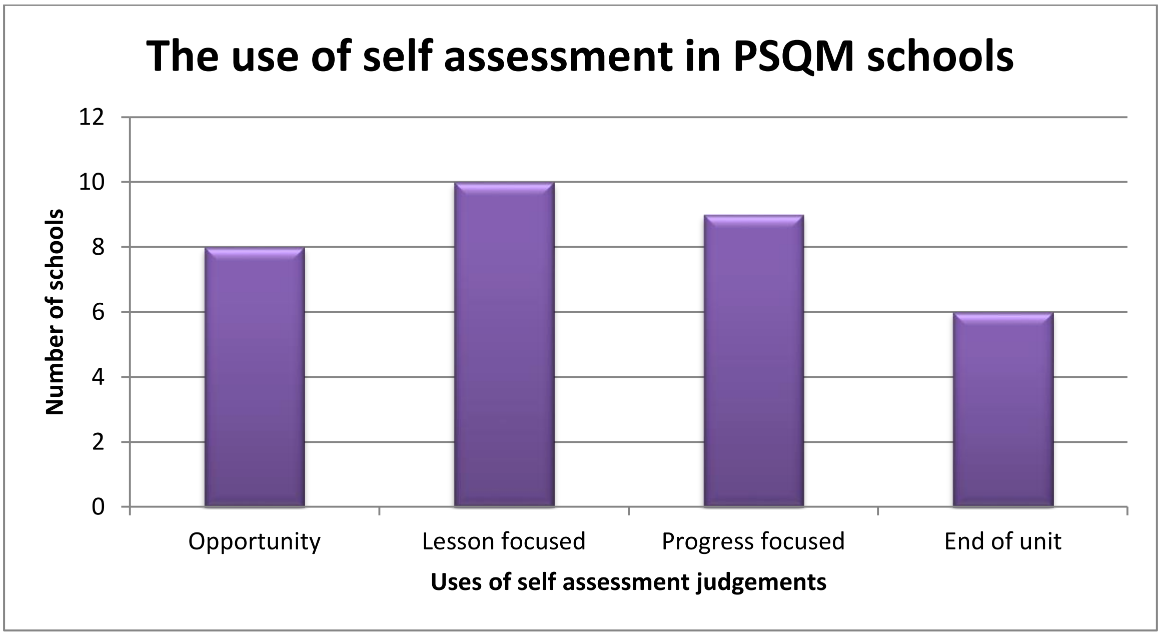

Figure 4: How Self Assessment was described by the 33 schools mentioning it in PSQM Round 4 (March 2013)