Earle, S. and Turner, J. (2023) 'What has happened to teacher assessment of science in English primary schools? Revisiting evidence from the Primary Science Quality Mark', *Research in Science & Technological Education*, 41 (1), pp. 22-38.

What has happened to teacher assessment of science in English primary schools? Revisiting evidence from the Primary Science Quality Mark

Dr Sarah Earle[a]* and Jane Turner[b]

[a]*Bath Spa University, Bath, UK*

[b]*University of Hertfordshire, Hatfield, UK*

*corresponding author

Email: s.earle@bathspa.ac.uk

Twitter: @PriSciEarle

ORCID: https://orcid.org/0000-0002-6155-1098

What has happened to teacher assessment of science in English primary schools? Revisiting evidence from the Primary Science Quality Mark

Abstract

Background: An earlier article in this journal (Earle 2014) provided a mapping of primary science teacher assessment practices in England using teacher reflections written for Round 4 of the Primary Science Quality Mark (PSQM).

Purpose: In the intervening years, the National Curriculum and statutory assessment system of levels has been replaced with a system of age-related expectations and a call for schools to develop their own systems of assessment. The aim of this study is to find out whether schools have changed their practices and to consider the ongoing changes they are making to their systems of assessment.

Sample: The previous analysis of 91 Round 4 schools in March 2013 was compared to teacher reflections from 200 schools from across England in the more recent PSQM Round 13, June 2017.

Design and methods: Qualitative content analysis was used to code the reported assessment methods in PSQM Round 13. The frequency of assessment methods were compared to the Round 4 mapping.

Results: Reported practice in summative assessment in Round 13 included more 'ongoing' assessment information, which could potentially support a more valid sampling of the construct, but there was a lack of clarity regarding the use of data tracking systems. In the PSQM year, subject leaders described starting or developing the use of a range of published and school developed resources to support assessment practices in formative and summative assessment.

Conclusion: Changing statutory assessment processes instigates change in school practices, but without support for teacher assessment literacy, the implementation of new assessment systems may not lead to clear processes for assessment which support learning in primary science.

Keywords: teacher assessment, primary science, teacher assessment literacy, summative assessment

Introduction

Assessment continues to be problematic for schools, with contrasting views at all levels of education regarding its purpose, implementation and impact. In line with many other countries (e.g. Finland, Australia, Scotland and Wales) formative and summative assessment of attainment in primary science in England relies on teacher judgement. Exactly how teachers are making such judgements was the focus for a previous article in this journal (Earle 2014), which provided a mapping of teacher assessment practices across a sample of schools in England in 2013. However, there have been dramatic changes in the assessment landscape since then, which will be discussed fully below, meaning that a new mapping of teacher assessment practices is required, in order to both analyse the impact of the changes and to support schools with this complex issue.

The Primary Science Quality Mark (PSQM) is an award scheme which supports primary schools to develop science leadership, teaching and learning (White et al. 2016). It requires a teacher with responsibility for science, the science subject leader (SL), 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 PSQM database was analysed in order to map primary science assessment practices in English schools at March 2013 (PSQM Round 4, Earle 2014). The Round 4 study found that formative and summative assessment were predominantly described separately, and that schools were largely using a combination of methods to make summative judgements, for example, tests for conceptual understanding and tracking grids for inquiry skills. This study seeks to look again at the national context, through analysis of Round 13 PSQM school submissions (June 2017), to consider how assessment practices have changed over time.

Teacher Assessment

Teacher assessment has been proposed to be a more valid way of assessing attainment (Gardner et al. 2010) because it can draw on a wider range of information than a standard summative test, providing a better sampling of the construct (Stobart 2009). However, concerns remain regarding the reliability of teacher assessment (Johnson 2013), since drawing on a wide range of information is both difficult to verify with others who have not taken part in the same classroom experiences, and open to bias, with teacher expectations affecting pupil tasks and outcomes (Campbell 2015). Nevertheless, if it is recognised that no assessment can be perfectly valid and reliable (Harlen 2007), since they pull in different directions, with validity improving with wider sampling and reliability improving with narrower sampling; it is about aiming for 'good enough' for the purpose, a balancing in practice (Earle 2017). The reliability of teacher assessment can be improved by clear criteria, moderation and training (Harlen 2009).

Teacher assessment can be used formatively, to support pupil learning (Assessment for Learning, AfL, Black and Wiliam, 1998), and summatively, to judge attainment at a particular point in time. It is the purpose or use of the assessment which denotes whether it is formative or summative, not something inherent in the assessment strategy itself. 'What is distinctive about assessment for learning is not the form of the information or the circumstances in which it is generated, but the positive effect it has for the learner' (Klenowski 2009: 264). Thus the teacher's role is integral, selecting and implementing assessment strategies and using the information to support pupil learning and/or report on attainment. Effective use of classroom assessment strategies does rely on teacher assessment literacy, with the teacher needing to understand and apply assessment principles and processes (DeLuca and Johnson 2017), which is a concern since assessment has been found to be the weakest aspect of teacher practice (Black and Harrison 2010).

A group of experts, convened by the Nuffield Foundation, proposed a pyramid-shaped model of 'formative to summative' assessment whereby the classroom practices would feed up the pyramid into later summary reports of attainment (Nuffield 2012). The Teacher Assessment in Primary Science (TAPS) project operationalised and exemplified this model into a school self-evaluation framework (Davies et al. 2017, Earle et al. 2017). Mansell et al. (2009) warn against using assessment information for multiple purposes, however, the negative impact they describe is seen when assessment data is used for institutional monitoring rather than for supporting pupil learning (p8).

Thus a ‘formative to summative’ model can be applied, if the primary focus remains on pupil learning rather than becoming a ‘tick-box culture’ (Mansell et al. 2009: 22), which again relies on the assessment literacy of the teachers when implementing new processes.

Assessment in England

Primary teachers in England have a statutory requirement to summatively assess each child against the National Curriculum descriptors in English, mathematics and science at ages 7 and 11 (DfE 2013a, STA 2017a). Standard Attainment Tests (SATs) for science for 11 year olds in England were removed in 2009; although testing has continued for English and maths and is used as the basis to measure school performance. Between 2009 and 2015 summative teacher assessment consisted of ascertaining a level for each pupil in science, continuing the system introduced in the Task Group on Assessment and Testing (TGAT) report (DES 1988). Whilst many teachers did not regret the removal of science SATs, the subsequent increased emphasis on making reliable teacher assessment judgements has caused concern (Turner et al. 2013: 3) and there were further concerns over perceived reduced status of primary science due to its lack of alignment with English and mathematics, as discussed further below.

Since the last study (Earle 2014), the TGAT ‘levels’ structure for assessment was removed and replaced by a system based on age-related expectations. The move from level descriptors to age-related judgements was seen as a radical shift for schools (Commission on Assessment without Levels 2015). After using level descriptors for more than 20 years, there were suggestions that the system was leading to the unhelpful labelling of children and teaching to the ‘test’ since schools were held accountable for results which were published in school performance tables. In addition, there had been a change in perception of the TGAT level 4, which had begun as a pupil average, but had become a target for all (Whetton 2009). The expectation at the time of writing is that by the end of the Key Stage (age 7 and 11), *“pupils are expected to know, apply and understand the matters, skills and processes specified in the relevant programme of study”* (DfE 2013a: 4), with the curriculum objectives becoming the new criterion scale. Thus the continuum of broad level descriptors has been replaced by more narrow and numerous criteria directly linked to age.

The new National Curriculum (DfE 2013a) for Key Stage 1 (ages 5-7) and Key Stage 2 (ages 7-11) was introduced in September 2014. The curriculum set out a year-by-year programme of study for science, organised into ‘Working Scientifically’ (scientific inquiry) and topics of biology, chemistry and physics such as: plants, everyday materials and electricity. Guidance explicitly stated that Working Scientifically must not be taught as a separate strand, *“but must always be taught through and clearly related to the teaching of substantive science content in the programme of study”* (DfE 2013a: 5). In the summer of 2015, children in Year 2 (age 7) and Year 6 (age 11) were the last to receive an end-of-key-stage ‘level’. Schools were encouraged to create their own assessment systems (DfE 2017b). However, the Commission on Assessment without Levels (2015) noted that: *“the system has been so conditioned by levels that there is considerable challenge in moving away from them…[with] some schools are trying to recreate levels based on the new national curriculum”* (p4), for example, creating new systems of ‘emerging, expected, exceeding’.

Any new assessment arrangements may take several years to become an established feature of classroom practice. Many note time as an important factor, since change in assessment practice: *‘requires regular and sustained opportunities for professional dialogue’* (Black and Harrison 2010: 207). Webb and Jones (2009) found that development of assessment practices, from *‘trialling’* to *‘integrating’* to *‘embedded’*, required not only changes in teacher values but also change in classroom culture, which is both difficult and takes time. Black and Wiliam (1998) suggested that change in assessment practice is likely to be slow and individual, but they also described the importance of real examples to support such changes, suggesting that exemplification may be a way of supporting the development of teacher assessment in primary science.

Stobart (2009) suggests that teachers are more confident with their judgements at Key Stage 1 (age 7) because of the lower stakes of these assessments; teachers are trusted to make judgements because their results are not used in school performance tables. However, he suggests that the higher stakes context of Key Stage 2 would make: *"any teacher assessment suspect given the importance of good results to a school"* (Stobart 2009: 174), indicating either a pressure to inflate results or a need for what would be seen as more reliable numerical evidence in such a high stakes arena. Perhaps this leaves science in an enviable position compared to English and mathematics, since science currently does not feature in league table accountability measures. If science assessments are not high stakes, then it follows that there should be fewer issues with reliability of teacher assessment, provided guidance and moderation are in place. However, with the accompanying drop in status of science, the issue becomes one of time, both to teach science and for assessment training or moderation. It appears primary science is stuck between a rock and a hard place: it needs high stakes assessments to ensure status, but low stakes assessments to ensure reliable teacher assessment.

Status of primary science in England

The status of primary science directly impacts on the amount of curriculum time for pupils and development time for teachers. Whilst the removal of standardised testing in Wales arguably led to increased opportunities for investigative work (Collins et al. 2010), a survey from the Wellcome Trust (2011: 1) found that teachers in England reported: *"less teaching time devoted to science; change to the status of science; science assessments not done; reduced curriculum or coverage of the curriculum"*. However, the removal of science testing was not the only factor, since Boyle and Bragg (2005) had already found substantially reduced teaching time for science, which they suggested was due to national strategies focused on raising test scores in English and mathematics (p435). More recent reports have also noted a lowering of status in primary science in England, often suggesting that science in primary schools has been side-lined by a continued focus on English and mathematics (e.g. Ofsted 2013, Wellcome Trust 2014, CBI 2015). The reduced status of science has led to a limited amount of lesson time, for example, an hour per week or less in one third of schools surveyed by the CBI (2015). In addition, a recent survey commissioned by the Wellcome Trust found that 58% of classes were not receiving two hours of weekly science (CFE Research 2017). A key challenge for primary science is to secure sufficient weekly curriculum time, making manageability of assessment processes a key priority in the current climate.

Eady's (2008) study of the purpose of teaching science in primary school found that many teachers saw scientific knowledge as paramount for performance in end of Key Stage testing. If teachers saw a strong relationship between the purpose of primary science and the passing of tests, the removal of those tests in 2009 could be one of the reasons for the reduced status of primary science. In contrast, Stobart (2009) suggests that a narrow focus on outcomes and tests is counter-productive because whilst it appears to raise the status of science, it is at the expense of a broader curriculum and deeper learning (p176). Eady (2008) also suggested that the commonly used QCA schemes of work (DfEE/QCA 1998) provided a progression of pre-planned lessons that negated the need to elicit pupil ideas; with a change in National Curriculum (DfE 2013a) these QCA schemes also became obsolete. Thus there is perhaps a generation of teachers for whom primary science was seen as a body of knowledge, with a pre-defined order and progression to be delivered in line with the QCA scheme of work, which was to be revised then tested and levelled externally at the end of the Key Stage. In recent years, the tests, levels and QCA scheme have all been removed, leaving teachers lacking supportive statutory structures and perhaps an uncertainty regarding why and how to teachi primary science.

Concerns have also been raised regarding the support which teachers receive for the assessment of primary science (Ofsted 2013). It was recommended that schools should: *'provide subject-specific continuing professional development for subject leaders and teachers that improves the quality of assessment and feedback for pupils in science'* (Ofsted 2013: 7). It appears that there is a need for professional development and support for science assessment, but the low status of primary science may limit the amount of time and resources schools feel they are able to devote to this.

Methods

The study utilised two pre-existing datasets of PSQM reflections: 91 schools in Round 4 (March 2013) and 200 schools from Round 13 (June 2017). Each teacher reflection consisted of around 200-400 words and contained description of: practice within the school, changes across the PSQM year, their impact and possible next steps. One of the 13 PSQM criteria (C2) required the subject leader to explain how science was assessed within the school, so it was the C2 reflection that was analysed to provide a mapping of assessment approaches taken by English primary schools. The PSQM datasets provide a sample of schools from across England, however, this could not be considered a representative sample since the schools were self-selecting by way of their PSQM application, which could mean that their practices are different to other schools. They were working towards the Primary Science Quality Mark which required them to reflect upon, and perhaps develop, their assessment practices, so it is quite possible that non-sampled schools will have less developed assessment practices. In addition, the reported practice may have been presented in a positive light, in support of their award application. Nevertheless, sampling can be described as a balance between what is ideal and what is possible (Newby 2010) and whilst it is acknowledged that this is only a subset of 'interested' schools, the PSQM dataset enables a national sample to be collated.

The C2 PSQM criterion that subject leaders participating in R13 wrote reflections against is as follows: '*The purpose of science assessment is well understood and shared by members of the school community. Assessment approaches are designed to fit those purpose'* (PSQM Handbook 2016). The criterion is a development from an earlier iteration of the PSQM criteria (2009-11) where subject leaders were required to demonstrate that formative and summative strategies were used to assess science in their schools. The change was made as an attempt to elicit a more evaluative response from subject leaders.

The C2 reflection for each school was anonymised at the point of download from the PSQM server. All PSQM schools are informed at the point of upload that their submission may be used anonymously for research purposes. The R4 and R13 schools received an additional email, to comply with ethical procedures (BERA 2018), providing them with the option to withdraw their data from the study.

Bearing in mind the significant changes to statutory summative assessment during this period, this study focused on the following research questions (RQs) to explore changes over time in assessment practice:

RQ1. How did schools in this sample make summative judgements of primary science?

RQ2. How have summative assessment practices changed between Round 4 (2013) and Round 13 (2017)?

RQ3. How have teachers in this sample developed their assessment practice during their PSQM year?

Qualitative content analysis was undertaken on the PSQM reflections (Silverman 2011), with coding supported by ATLAS.ti to allow for 'constant comparison' of codes and quotations to rigorously cross-reference and refine coding (Robson 2011, Coe 2012). The codes were not fixed at the point of inception, they were open to change as more data was examined (Simons 2009), which meant that early coded items were re-visited a number of times to ensure consistency across the dataset.

The R13 analysis for RQ1 and RQ2 began with theory-led coding, based on the previous R4 findings, with emergent codes added from the data. The focus for this study was on summative assessment methods. To allow for direct comparison between R4 and R13, some of the categories from R4 (Earle 2014) were revisited, for example, the R4 'levels on planning' became 'criteria on planning' since the levelling system had been removed, as discussed above. The R13 analysis for RQ3 followed the same method but coded for any actions described as developments to assessment practices, and included both formative and summative strategies.

The coding of data enabled a numerical analysis of the qualitative data, to complement the prose and provide a survey of the whole dataset (Silverman 2011: 379). Using quantification in qualitative data analysis is not the same thing as adopting a quantitative methodology; the data was not a measurement of practice (Bryman 2012: 35). Quantification of the data did not rely on key word frequencies, since teachers mentioned strategies in a range of ways, for example, sometimes proposing next steps rather than listing current practice or recent changes to practice. The contextual nature of the data required reading and re-reading to extract the use of the assessment techniques. In this research, numerical summaries were used to support analysis of the prevalence of an assessment strategy. One final point to note is that the word 'tracking' is used with the UK meaning of recording of data, rather than the US meaning of organising pupils into groups, which in the UK is called 'setting' and rarely used in primary science.

Results

RQ1. How did schools in this sample make summative judgements of primary science?

In order to answer RQ1 a practical definition of 'summative' which could be applied consistently to this dataset was constructed. The approach was classified as summative if:

- it was described as 'end of unit' or 'end of year'.
- it fulfilled a summarising purpose, e.g. reporting to next teacher or put into the school tracking software (where a 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'.

A range of summative methods were identified, which are briefly defined and exemplified in Table 1.

Table 1: Definitions and examples of summative methods in Round 13

Code	What information feeds into summative judgements?	Example from dataset
Tests primarily	tests are described as main evidence for judgement	‘All children are assessed at the end of each unit of work (approximately each half term) in a written test. The test asks the children to recall subject knowledge and asks them to devise a test, or element of a test in order to assess their experimental understanding. These tests are then given a percentage and a grade generated.’ R13.36.2.S
Tests + other	tests described as key part of information feeding into judgements, supplemented by other evidence	‘This led to the task of me creating an assessment booklet, meaning that staff would have access to the end of unit tests and self-evaluation sheets for all year groups. They are based on our scheme and link with the resources we use from G. All year groups now have and complete an end of unit written assessment and the evaluation/tracking sheet for each topic.’ R13.13.2.S
Tests + tracking	data ‘tracking' as an assessment tool, also use tests	‘I have introduced Science Assessment sheets throughout the School, which are based on the Curriculum Framework and tailored to each Year group. The staff have been using these Assessment sheets since February of this Year and are able to track the progress of their class, as individuals. Each class has carried out knowledge based tests in Science and these will continue to happen, going forward.’ R13.15.2.S
Tracking + other	data tracking sheet/software as an asst tool, plus other e.g. self asst	‘All staff are now confident in using the L document for assessment including the skills ladders and they are all using an assessment recording proforma that I have developed (slide 11). Although these documents have been useful in helping teachers to make judgements, I have also been trying out a variety of assessment techniques during science sessions in my class which challenge the children to use the knowledge and skills they have developed during a particular science topic.’ R13.26.2.B
Tracking grids primarily	ongoing data tracking sheet/software described as main method of asst	‘I have set up topic assessment grids that provide the teacher targets to help base their planning on and a means to assess a pupil against to make a summative judgement at topic end.’ R13.32.5.B

‘Ongoing’ range	lists range of strategies/evidence types which are collected over time and feed into judgement, opportunities could be in any lesson	‘I introduced formative strategies to be used school wide including KWL grids for pre and post unit assessment, hinge point questions to check understanding at key areas of misconception and graphic organisers to unpick understanding in a way that is usable for all ages (CPD log/staff meeting minutes). Staff from all year groups then had a bank of evidence to draw upon to make formative decisions for planning and ultimately summative decisions at the end of a unit of work.’ R13.15.19.S
Regular tasks	one off tasks (not every lesson), programme of specific tasks, may be repeated e.g. pre/post	‘The assessment tasks clearly shows the level of knowledge at the beginning of a unit of work (cold task) and the learning that has then taken place during the sequence of lessons (hot task) in every year group of the school.’ R13.12.2.S
Criteria on planning	assessment against criteria for each lesson on planning	‘Foremost, on our planning, we have included criteria for children who are working at greater depth and those who are working towards it. Planning is annotated by the class teacher with the children who are working at the different levels which is returned to the subject leader.’ R13.5.7.S
End of unit other	teacher assessment specified as taking place at end of unit	‘Post-unit assessments are carried out at the end of each topic to measure progress and children working above, below and those who are on track are identified in each class.’ R13.15.10.G
Summative unclear	no clear description of summative assessment provided	‘During this year teachers have grown in confidence when assessing science and are making better use of the resources available to them in order to accurately identify a child’s attainment and agree their next steps.’ R13.37.2.S

The frequency of the reported summative methods across the sample is shown in Table 2 and Figure 1.

Table 2: Summative methods in Round 13 (N=200)

Summative methods	**Frequency in R13**
Tests primarily	7
Tests+other	18
Tests+tracking	6
Tracking+other	25
Tracking grids primarily	36

Ongoing range	55
Regular tasks	14
Criteria on planning	7
End of unit other	28
Summative unclear	4
Total	200

Figure 1: Summative methods in Round 13 (N=200)

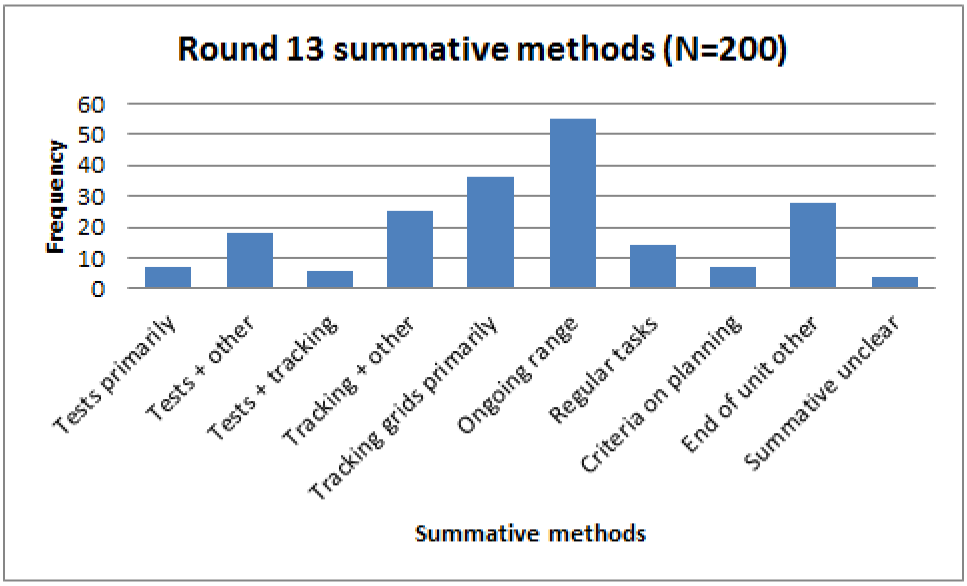


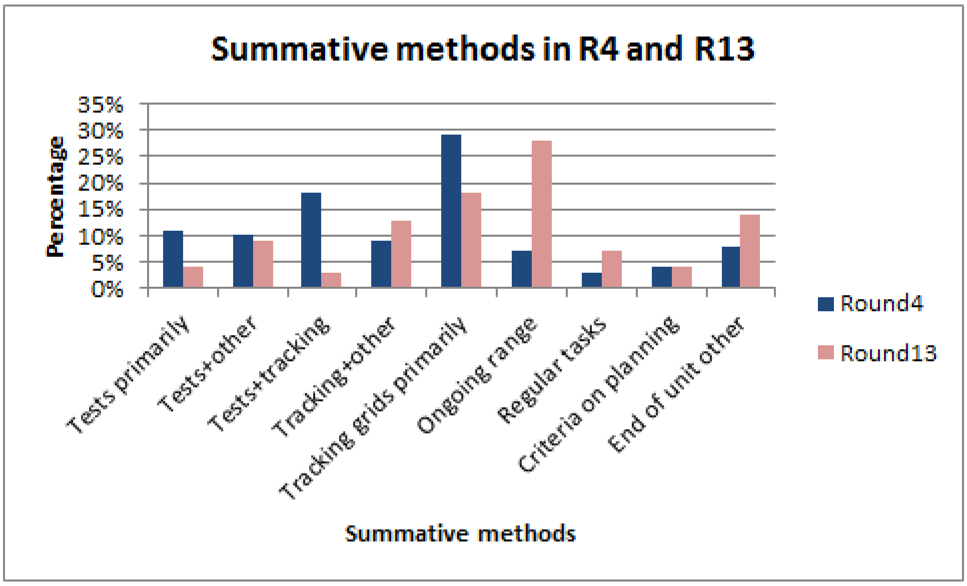
Table 2 and Figure 1 illustrate that the assessment methods used by English primary schools to make summative judgements were wide-ranging. It is important to note that the majority of summative methods in the R13 PSQM submissions describe methods which draw on assessment information gathered over time: ‘tracking’, ‘ongoing range’, ‘regular tasks’ and ‘criteria on planning’. The implications for this with regard to validity will be considered in the Discussion below.

The examples in Table 1 provide an insight into the range of practices described by the subject leaders in the sample. The descriptions of summative methods varied widely in their clarity, but the most problematic was the use of ‘tracking’. Many of the subject leaders were not clear about whether ‘tracking’ was used to record the results of assessments, or whether ‘tracking’ was seen as a means of assessment itself. This issue will also be discussed further below.

RQ2. How have assessment practices changed between Round 4 (2013) and Round 13 (2017)?

A comparison between the summative assessment methods described in R4 (March 2013) and R13 (June 2017) is presented in Figure 2. 'Summative unclear' schools are not included (R4=2, R13=4).

Figure 2: Summative methods in Round 4 (N=89) and Round 13 (N=196)



The key differences between R4 (March 2013) and R13 (June 2017) are listed below, then further explored in the discussion section:

- A new National Curriculum structure is in place: levels have been replaced by age-related expectations.
- The use of tests is down, with 39% of R4 schools mentioning tests and only 16% of R13 schools, perhaps due to the lack of products related to the newer curriculum.
- In R4, 18% of schools described using tests for concepts and tracking for skills, but this was only recorded for 3% of the R13 schools.
- Although there had been a drop in 'tracking grids primarily' from 29% in R4 to 18% in R13 (perhaps due to the high use of 'Assessing Pupil Progress (APP) grids' (DCSF 2010) in R4 which had been introduced by a previous government as non-statutory guidance, but no

longer matched the curriculum), this figure is still high this figure is still high.As noted above, the recording of assessment data was assumed to be a method of assessment for many.

- A big increase in schools describing a range of ongoing activities feeding into summative judgements: from 7% or R4 schools to 28% of R13 schools.

RQ3. How have teachers in this sample developed their assessment practice during their PSQM year?

A practical definition of 'assessment practice', which could be applied consistently to this dataset, was constructed. The practice was classified as an as assessment practice if it described:

- strategies used to elicit children's understanding and skills in science
- strategies used to make judgements of children's attainment in science
- strategies used to track children's progress in science

A range of developments in assessment practice were identified. These new and supplemented practices are briefly defined and exemplified in Table 3. Practices coded in less than 5% of the data set (portfolios and pupil targets) are not included in the results below.

Table 3: Definition and examples of changing Round 13 assessment practices

Code	What developments in practice have been made?	Example from dataset
Started using a published scheme of work which includes assessment	A scheme of work for science which includes an assessment resource has been purchased and implemented by the school	'The school appreciates that testing is not the only means to assess, therefore we purchased the SNAP Science assessment that will be trialled throughout the summer months once resources are prepared for each year group.' R13.22.1.G

Started or increased use of tests	The use of commercially published written tests has been introduced or increased.	‘As well as the summative assessment that gives an overall level of skills that can be shared with pupils and parents there are six formal written tests in Year 3,4 and 5 and three in Year 6. These are put together using Testbase and are bespoke tests that closely mirror the practical experiments and learning that has gone on in that unit.’ R13.12.3.S
Started or increased use of published assessment resources	The use of published resources to elicit children’s understanding has been introduced or increased.	‘Identified need to source assessment tools to support how we make informed accurate assessments our pupils. The subject leader, Concept Cartoons and Snap Science Snap Shot Assessment Tasks to facilitate this.’ R13.12.1.S
Increased range of AFL strategies	The range of pedagogic strategies used to elicit children’s understanding has increased.	‘We discussed a broad range of activities that you could do in the classroom to assess progress of children. These included prior and post learning tasks and questionnaires. Questioning has become a key element of lessons for teachers to assess the understanding of pupils. They are able to include various AfL strategies into their lessons and assess on how they can adapt their planning for future lessons.’ R13.17.4.S
Started using school developed tracking system	A bespoke system for tracking children’s attainment has been developed and implemented in school	‘Science is assessed termly using the whole school assessment tool called Learning Ladders. The system is a bespoke made system, where the science criteria has been developed by the science leaders, after consultation with staff, in line with the NC. The proposed science criteria for assessment were discussed, modified and agreed upon at a staff meeting. They came into effect in term 6. The assessment criteria include knowledge and investigation skills. Teachers assess the children’s learning and once a term, they are required to upload the data onto the school system in the form of a tick.’ R13.22.5.S

Started using commercial tracking system	A commercial system for tracking children’s attainment has been implemented in school	‘During the same staff meeting I proposed the idea of using Target Tracker to give teachers the opportunity to assess their class throughout the year against a number of key objectives (we currently use the programme to assess in Reading, Writing and Mathematics). We agreed to initially use the programme to check we are covering the objectives in our lessons and our appropriate next step for this would be to begin using it on a trial basis so all staff members can become confident in using it and then we can fully assess the effectiveness of it next year.’ R13.28.1.S
Started using school developed in school periodic assessment activities	Bespoke end of unit of end of topic assessment activities have been developed and used in school	‘A practical assessment task is carried out each term as well as a number of knowledge assessment tasks, which were created alongside two other schools in the area to track progress.’ R13.17.4.S
Starting using school developed progression criteria	A set of statements indicating progression criteria against attainment objectives has been developed and is used in school.	‘At the beginning of the year each year group’s curriculum coverage was linked to ‘I can’ statements. These were then distributed to staff so progress could be checked throughout the year ensuring full curriculum coverage and allowing staff to gather a clear picture of children’s progress.’ R13.30.4.S
Moderation of science assessment judgements	Teachers have compared evidence used to make attainment judgements to moderate assessment.	‘We have looked through the books and worked together to assess the children to make sure we agree that the children correctly assessed. We have also moderated our assessments with another school.’ R13.3.14.G
Started using exemplification materials	The use of published materials which exemplify standards have been used to check teachers’ assessment judgements.	‘I also highlighted the exemplification materials provided by the government to ensure judgments were moderated and robust.’ R13.28.4.S
Lesson plans changed to fit new assessment approaches	School lesson planning models have been adapted in response to changes	‘Following this meeting, teachers have aimed to include different methods of assessment in their planning, and the SL has given support with this by team planning.’ R13.4.3.G

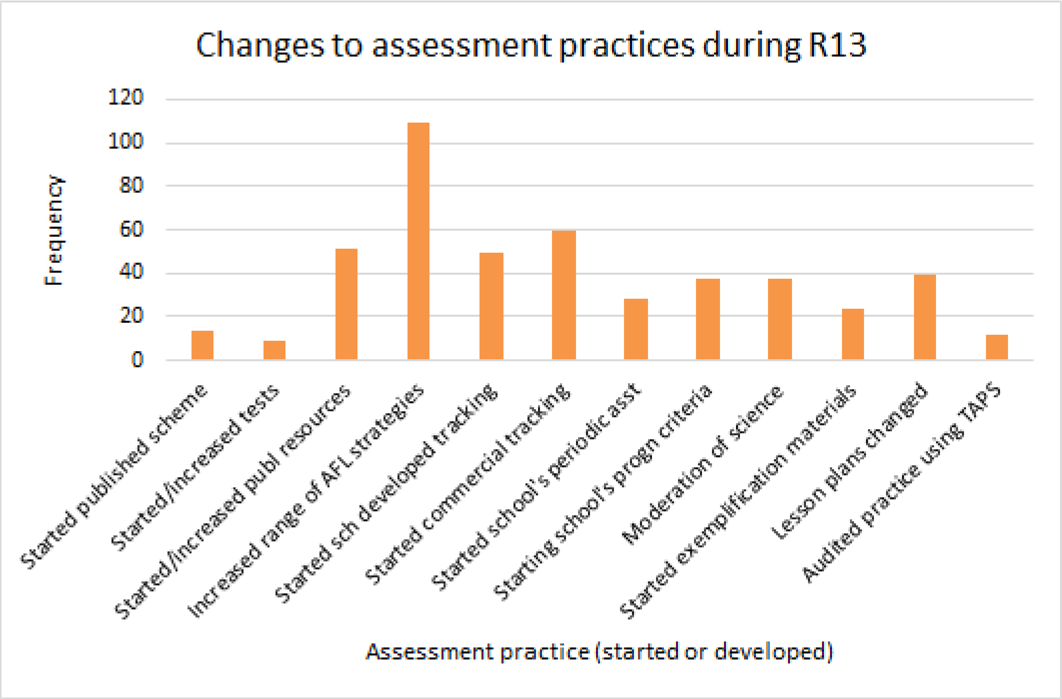
	to assessment practices.	
Audited practice using TAPS pyramid	Used the TAPS pyramid to audit current practice in assessment in science.	‘I have introduced the TAPS pyramid tool to provide teachers with a supportive structure to evaluate and develop their assessment processes.’ R13.9.1.S

Having defined and exemplified the changes made to assessment practices identified in the R13 descriptions, we will now move on to explore the frequency of such strategies within the dataset. A tally of the reported assessment techniques across the sample is shown in Table 4 and Figure 3.

Table 4: Changes to assessment practices in Round 13

Assessment practice	**Number of R13 schools starting or developing this practice (N=200)**
Started published scheme	13
Started/increased tests	9
Started/increased published resources	51
Increased range of AFL strategies	109
Started using school developed tracking system	49
Started using commercial tracking system	60
Started using school developed periodic assessment activities	28
Starting using school developed progression criteria	37
Moderation of science assessment judgements	37
Started using exemplification materials	24
Lesson plans changed	39
Audited practice using TAPS pyramid	12

Figure 3. Changes to assessment practices in R13 (N=200)



All subject leaders reported at least one aspect of assessment practice where there had been development and most reported several. Participation in the PSQM process requires subject leaders to evaluate science assessment practice in their schools and so it should be expected that development occurs during this year. Changes to assessment practices take time, with all teachers identifying next steps for future developments. There was a clear recognition that the statutory requirements had changed and that assessment of science in schools had required review and development in line with this. The data shows that they did this in different ways, including purchasing new commercial resources and developing their own, to support developments in both formative and summative assessment.

Discussion

The data presented above indicates both that there have been major changes to the landscape of primary science assessment practice in the samples of schools between R4 and R13 (RQ1&2), and that the R13 schools have made changes to assessment practice within the PSQM year (RQ3). It is important to remember that this is a self-selecting sample of schools working towards an award and that PSQM provides a supportive structure for development, so this sample cannot be considered to be representative of all English primary schools. In addition, the assessment criterion was reworded between R4 and R13 to support teachers to consider the purposes of assessment, rather than listing formative and summative strategies, so teacher reflections from R13 may emphasise different parts of assessment practice to those from R4. However, the statutory structural shift from a system based on broad levels to a system based on detailed age-related expectations (DfE 2013a) is likely to have been the bigger driving force behind the described changes, with schools in the R13 sample clearly recognising the need to develop practice.

A recurring term used by schools in the data-set was 'tracking', which was problematic because it did not appear to have a consistent meaning or use for schools. Some described tracking as a tool for assessing, whilst others described tracking as a tool for recording of data, electronically or on paper. There is a concern that the 'datafication' (Roberts-Holmes and Bradbury 2016) of school accountability processes have raised the importance of data 'tracking' to the point that it is now seen to be the end goal for classroom assessment: tracking as assessment. However, in order to place a number into the tracker or click to say an objective has been achieved, there needs to have been a teacher judgement in advance. Ignoring the teacher role in completing the tracker risks a lack of attention to the process of teacher assessment and the importance of training for teacher assessment literacy to make these judgements valid and reliable. There is also a risk when the information is passed on to senior leaders, that spreadsheet data is taken to be objectively true (James et al. 2007: 385), without recognition of the processes behind the numbers and percentages. Tracking is a system of recording, which can be useful for the questions and actions it generates (Peacock 2016: 100); tracking is not necessarily a system of assessment, and future research should explore this area further.

Nevertheless, if the tracking grids, whether paper or electronic, provide clear and broken down criteria against which judgements could be made, then teacher understanding of the criteria could be developed. Embedding the National Curriculum objectives into planning as learning objectives and periodically making judgements regarding whether children are meeting these objectives, provides information which could be used formatively, as well as informing summative summaries. Thus the usefulness of tracking grids depends on whether the criteria are clear enough to provide support for a shared understanding of progression.

A key difference between the summative methods employed in R4 compared to R13 appears to be the increase in descriptions of more 'ongoing' assessment (e.g. 'ongoing range', 'regular tasks' and 'criteria on planning') and a corresponding drop in the use of end of topic tests (from 39% to 16%). As noted above in RQ1, changes to the curriculum could have been a major influence here, with a lack of published test material matching the new objectives. In addition, and perhaps more significantly, the move from 'best fit' broad levels to a longer list of age-related expectations could have meant that schools needed to make assessment judgements over a longer period of time.

There are a number of implications of a move from formal testing to ongoing assessment. Reduced testing could lead to cohorts of pupils who are inexperienced in formal science assessments, which is perhaps one of the reasons for the low scores in the biannual national sampling tests, with only 23% of 11 year olds achieving the expected standard (STA 2017b). Although, the low status of science discussed in the introduction is likely to have a more dramatic effect on the results of STA's national sampling. The reduction in explicit mention of tests in the R13 sample does not mean that schools are not using them, it could be that their use has changed: from sole measure to part of the information. Alternatively, teachers may have chosen not to write about tests in their reflections, for example, if they associated tests with the harmful effects of curriculum narrowing (Wiliam 2003), choosing instead to focus on the new formative strategies they had been trialling that year (which dominates Figure 3).

Using a broader range of assessment information from ‘ongoing’ assessment could enhance validity (Stobart 2009), since there is likely to be a wider sampling of the construct by collating and summarising assessment information from a range of activities. Such summative assessment could be termed ‘summary’ rather than ‘snapshot’ (Earle 2018), where a ‘summary’ draws on a number of sources, in contrast to a ‘snapshot’, which only provides information about limited content, at one point in time. If a ‘formative to summative’ model of assessment (Nuffield 2012, Davies et al. 2017) is to be followed, then the assessment information which is initially gathered for formative purposes, can inform a summative ‘summary’ judgement. The data presented in this article suggests that the separation of formative and summative methods seen in Earle (2014) is less pronounced, with the increase of ‘ongoing’ methods. This suggests that classroom activities are being used to provide assessment information, which can be used formatively to adapt lessons on a day-to-day basis, but can also be utilised to inform summaries for reporting purposes.

A question could be raised regarding the regularity of ‘ongoing’ assessment processes, for example, whether the assessment information should be gathered weekly or monthly. The detailed national curriculum criteria could lead to a constant assessment ‘tick box culture’ which Mansell et al. (2009) warned against, with surface-level performance taking priority over deeper learning. In addition, if teachers are trying to evidence attainment in all lessons, to ‘test’ at each encounter, then little time is left for teaching, consolidation and practice, with formative assessment misinterpreted as repeated ‘testing’ (Black and Harrison 2010). It is thus essential for the formative purpose to take precedence, whilst recognising that some of the information gathered might be useful when collating a summary. The desire to enhance validity requires a broader sampling of the curriculum, but this does not mean that every lesson will lead to information that can inform a summary judgement, especially near the beginning of a unit of work.

Assessment decisions, which are so intertwined with everyday teaching practice, require teacher assessment literacy. The changes to assessment practice made by the R13 schools (RQ3) required many to involve teachers in professional learning activities, for example, developing progression criteria, exemplification and moderation. Such activities build a shared understanding (Earle 2017), enhancing teacher assessment literacy by supporting both the reliability of teacher assessment (Harlen 2009) and the use a broader range of information when making assessment judgements. Consistency in teacher judgement can also be enhanced by utilising externally developed exemplification materials such as those provided by the STA (STA 2016), the Association for Science Education (www.ase.org.uk) and TAPS (www.pstt.org.uk). Use of exemplification and moderation are key to developing understanding of the attainment criteria, supporting both reliability of summative judgements and the process of formative assessment, since the teacher is better able to recognise progression in scientific understanding and skills. The authors propose that the development of teacher assessment literacy should be a priority for researchers, policymakers and schools, to ensure that assessment supports learning in primary science.

Wordcount: 7670 (including tables and references)

References

Association for Science Education (ASE) https://www.ase.org.uk/plan [Last accessed 9th September 2019]

BERA (2018) *Ethical guidelines for educational research* (4th Ed). London: BERA.

Black, P. and Harrison, C. (2010) Formative assessment in science. In J. Osborne and J. Dillon (Eds) *Good practice in science teaching: what research has to say.* Maidenhead: Open University Press.

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

Boyle, B. and Bragg, J. (2005) No science today – the demise of primary science, *The Curriculum Journal*, 16, 4: 423-437.

Bryman, A. (2012) *Social Research Methods*. Maidenhead: OUP.

Campbell, T. (2015) Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment, *Journal of Social Policy*, 44, pp 517-547

CBI (2015) *Tomorrow's World: inspiring primary scientists.* London: Confederation of British Industry.

CFE Research (2017) *State of the nation' report of UK primary science education: baseline research for the Wellcome Trust Primary Science Campaign.* Leicester: CFE Research.

Coe, R. (2012) Conducting your research. In Arthur, J., Waring, M., Coe, R. and Hedges, L. (Eds.) *Research methods and methodologies in education*. London: Sage.

Collins, S., Reiss, M. and Stobart, G. (2010) What happens when high-stakes testing stops? Teachers' perception of the impact of compulsory national testing in science of 11 year-olds in England and its abolition in Wales, *Assessment in Education: Principles, Policy and Practice*, 17, 3, 273-286.

Commission on Assessment without Levels (2015) *Final report of the Commission on Assessment without Levels.* London: DfE.

Davies, D., Earle, S., McMahon, K., Howe, A. and Collier, C. (2017) Development and exemplification of a model for Teacher Assessment in Primary Science, *International Journal of Science Education*, 39:14, 1869-1890.

DeLuca, C. and Johnson, S. (2017) Developing assessment capable teachers in this age of accountability, *Assessment in Education: Principles, Policy & Practice*, 24, 2, 121-126.

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

Department for Education (DfE) (2013a) *National Curriculum in England: science programmes of study.* London: DfE.

Department for Education (DfE) (2013b) Assessing without levels statement. Accessed June 14. http://www.education.gov.uk/schools/teachingandlearning/curriculum/nationalcurriculum2014/a00225864/assessing-without-levels

Department for Education and Employment (DfEE)/QCA (1998) *Qualifications and Curriculum Authority Schemes of Work.* London: DfEE/QCA.

Department of Education and Science (1988*) National Curriculum Task Group on Assessment and Testing (TGAT): A report.* London: Department of Education and Science and the Welsh Office.

Eady, S. (2008) What is the purpose of learning science? An analysis of policy and practice in the primary school, *British Journal of Educational Studies*, 56, 1: 4-19.

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

Earle, S. (2017) The challenge of balancing key principles in teacher assessment, *Journal of Emergent* Science, 12: 41-47.

Earle, S., McMahon, K., Collier, C., Howe, A. and Davies, D. (2017) *The Teacher Assessment in Primary Science (TAPS) school self-evaluation tool.* Bristol: Primary Science Teaching Trust.

Earle, S. (2018) *The relationship between formative and summative teacher assessment of primary science in England.* PhD thesis, Bath Spa University.

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

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

Harlen, W. (2009) Improving assessment of learning and for learning, *Education 3–13*, 37:3, 247-257.

James, M., McCormick, R., Black, P., Carmichael, P., Drummond, M., Fox A., MacBeath J., Marshall B., Pedder D., Procter, R., Swaffield S., Swann J. and William D. (2007) *Improving Learning How to Learn*. Abingdon: Routledge.

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

Klenowski, V. (2009) Assessment for Learning revisited: an Asia-Pacific perspective, *Assessment in Education: Principles, Policy & Practice*, 16:3, 263-268.

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

Newby, P. (2010) *Research Methods for Education*. Harlow: Pearson.

Ofsted (2013) *Maintaining Curiosity: A survey into science education in schools*. Report No. 130135. Manchester: Ofsted.

Peacock, A. (2016) *Assessment for Learning without Limits.* London: McGraw-Hill Education.

Roberts-Holmes, G. and Bradbury, A. (2016) Governance, accountability and the datafication of early years education in England, *British Educational Research Journal*, 42, 4, 600-613.

Robson, C. (2011) *Real world research*. 3rd edition. Chichester: Wiley.

Silverman, D. (2011) *Interpreting qualitative data: a guide to principles of qualitative research*. 4th edition. London: Sage.

Simons, H. (2009) *Case study research in practice.* London: Sage.

Standards and Testing Agency (2016) *2016 teacher assessment exemplification: end of key stage 2 Science.* London: STA.

Standards and Testing Agency (2017a) *Teacher assessment frameworks at the end of key stage 2: For use in the 2017 to 2018 academic year.* London: STA.

Standards and Testing Agency (2017b) *Key stage 2 science sampling 2016: Methodology note and outcomes.* London: STA.

Stobart, G. (2009) Determining validity in national curriculum assessments, *Educational Research*, 51, 2, 161–179.

Teacher Assessment in Primary Science (TAPS) project https://pstt.org.uk/resources/curriculum-materials/assessment [Last accessed 9th September 2019]

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.

University of Hertfordshire (2017) *Primary Science Quality Mark Handbook*. Hatfield: University of Hertfordshire.

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

Wellcome Trust (2011) *Primary Science Survey Report.* London: Wellcome Trust.

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

Whetton, C. (2009). 'A brief history of a testing time: national curriculum assessment in England 1989–2008', *Educational Research*, 51, 2, 137–159.

White E., Dickerson C., Mackintosh J. and Levy, R. (2016) *Evaluation of the Primary Science Quality Mark programme - 2013-15.* Hatfield: University of Hertfordshire.