



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

ResearchSPAce

<http://researchspace.bathspa.ac.uk/>

Your access and use of this document is based on your acceptance of the ResearchSPAce Metadata and Data Policies, as well as applicable law:-

<https://researchspace.bathspa.ac.uk/policies.html>

Unless you accept the terms of these Policies in full, you do not have permission to download this document.

This cover sheet may not be removed from the document.

Please scroll down to view the document.

The relationship between formative and summative teacher assessment of primary science in England

Sarah Earle

A thesis submitted in partial fulfilment of the
requirements of Bath Spa University
for the degree of Doctor of Philosophy

Institute for Education, Bath Spa University

March 2018

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement. I have exercised reasonable care to ensure that my thesis is original, and does not to the best of my knowledge break any UK law or infringe any third party's copyright or other intellectual property right.

ABSTRACT

Assessment drives the taught curriculum, defines what is valued (Stobart 2008) and can enhance or hinder learning (Mansell et al. 2009). In England, the complexities of assessment are compounded by ongoing changes to statutory assessment procedures and a lack of centralised guidance for judgements of primary science (Turner et al. 2013). The Nuffield expert group (2012) proposed a pyramid-shaped model of teacher assessment which utilised formative information to inform summative judgements. The model was operationalised by the Teacher Assessment in Primary Science (TAPS) project (Davies et al. 2014), but there was no explicit explanation of the ‘formative to summative’ process.

This study sought to develop understanding of the relationship between formative and summative teacher assessment of primary science, in action and over time. A Design-Based Research (DBR) approach was used in order to develop guidance directly relevant to practice. Analysis of 91 submissions from the Primary Science Quality Mark (PSQM) database provided a mapping of current practice in England. Two case studies of TAPS project schools considered the enacted relationship between formative and summative assessment, during implementation of a ‘formative to summative’ approach.

New insights have been gained into the conceptualisation and enactment of the relationship between formative and summative assessment. Teacher conceptualisations of assessment were found to encapsulate a wide range of dimensions such as: purpose, value, audience, assessor, timing, formality, rigidity and support. Refinements to the TAPS pyramid model are proposed to explain the ‘formative to summative’ process, conceptualising summative assessment as a summary judgement of attainment, which may be informed by snapshot and focused assessment activities. It was found that implementation of a ‘formative to summative’ approach required a shared understanding of key assessment concepts like validity and reliability; a seesaw balance model is proposed to support the development of such a shared understanding.

Contents

List of figures	11
-----------------------	----

List of tables	12
----------------------	----

Chapter 1 Introduction

1.1 Introduction to key concepts	13
---	-----------

1.2 Rationale and aim	15
------------------------------------	-----------

1.2.1 Rationale	15
-----------------------	----

1.2.2 Aim and research questions	17
--	----

Aim	17
-----------	----

Research questions	18
--------------------------	----

1.3 Context	18
--------------------------	-----------

1.3.1 Assessment in England.....	18
----------------------------------	----

1.3.2 Status of primary science in England	20
--	----

1.3.3 My own professional context and experience	22
--	----

1.4 Links between PhD and key projects	22
---	-----------

1.4.1 Key projects and organisations.....	22
---	----

1.4.2 Relationship between TAPS and PhD	23
---	----

1.5 Overview of thesis	25
-------------------------------------	-----------

Chapter 2 Literature Review

2.1 Introduction	27
-------------------------------	-----------

2.2 Validity and reliability in teacher assessment	28
---	-----------

2.2.1 Validity	28
----------------------	----

2.2.2 Reliability.....	30
------------------------	----

2.2.3 Validity and reliability in teacher assessment	32
--	----

2.2.4 Relationship between validity and reliability in teacher assessment	35
---	----

2.3 Formative and summative assessment.....	36
--	-----------

2.3.1 Distinctions between formative and summative assessment.....	36
--	----

2.3.2 A 'formative to summative' approach to teacher assessment.....	39
--	----

2.3.3 A model of 'formative to summative' teacher assessment.....	41
---	----

2.4 The nature of primary science	46
--	-----------

2.4.1 Purposes and nature of science education.....	46
2.4.2 Inquiry terminology	48
2.4.3 Atomism and holism in the process-content debate	49
2.4.4 Professional learning for teacher assessment literacy and subject leadership	53
2.5 Summary and research questions.....	55

Chapter 3 Methodology

3.1 Introduction.....	58
3.2 Design-Based Research.....	59
3.2.1 The nature of Design-Based Research in this study.....	59
3.2.2 Features of Design-Based Research.....	61
3.2.3 Phases of Design-Based Research.....	62
3.3 Research sample.....	64
3.3.1 Sampling	64
3.3.2 Case study within DBR.....	65
3.3.3 Case studies in this research	67
3.4 Methods of data collection.....	69
3.4.1 Overview of data collection methods	69
3.4.2 Documentary extracts from PSQM database.....	69
3.4.3 Documentation and written tasks within case studies.....	69
3.4.4 Observations within case studies.....	71
3.4.5 Discussions and interviews within case studies.....	72
3.5 Validity and reliability in social science research.....	75
3.5.1 Validity and trustworthiness	75
3.5.2 Internal validity and credibility	76
3.5.3 External validity and transferability	76
3.5.4 Consequential validity and authenticity	77
3.5.5 Reliability and dependability	78
3.5.6 Triangulation	79
3.5.7 Respondent validation or member checking.....	80
3.6 Ethics	81

3.6.1 Key principles	81
3.6.2 Roles of researcher	83
3.6.3 Roles of participants	84
3.7 Data analysis	85
3.7.1 Approach to data analysis	85
3.7.2 Theory-led and emergent coding.....	85
3.7.3 The process of coding	87
3.7.4 From codes to themes.....	89
3.8 Summary	91

Chapter 4 Formative and summative assessment in submissions to the Primary Science Quality Mark

4.1 Introduction.....	92
4.2 Summative assessment.....	93
4.2.1 Categorising as summative.....	93
4.2.2 Methods used by schools for summative assessment.....	95
4.2.3 The use of APP tracking grids	99
4.2.4 Summative assessment summary.....	102
4.3 Formative assessment	103
4.3.1 Categorising as formative assessment	103
4.3.2 Strategies for elicitation	104
4.3.3 Pupil involvement in assessment	109
4.3.4 Formative assessment summary	110
4.4 The relationship between formative and summative assessment.....	110
4.4.1 Separate descriptions of formative and summative	110
4.4.2 Links between formative and summative assessment	112
4.4.3 Separate systems for inquiry skills and conceptual understanding.....	115
4.4.4 Relationship between formative and summative summary	116
4.5 Summary and conclusions.....	116

Chapter 5 Case Study A: The relationship between formative and summative assessment

5.1 Introduction	118
5.1.1 Chapter overview	118
5.1.2 School A context	119
5.1.3 School A data and analysis	120
5.2.1 Whole school processes	121
5.2.2 Summary of practice at whole school processes layer	125
5.3 Summative reporting layer	125
5.3.1 Summative as ‘best fit’	125
5.3.2 A separate system for inquiry skills	129
5.3.3 Summary of practice at summative reporting layer	132
5.4 Monitoring layer	133
5.4.1 School structures.....	133
5.4.2 Moderation	136
5.4.3 Summary of practice at monitoring layer	139
5.5 Teacher layer	140
5.5.1 Teacher questions	140
5.5.2 Discussion and use of criteria in lessons.....	141
5.5.3 Recording and marking	144
5.5.4 Summary of practice at teacher layer.....	147
5.6 Pupil layer	147
5.6.1 Self and peer assessment.....	147
5.6.2 Summary of practice at pupil layer.....	149
5.7 Conclusion	150
5.7.1 Key features of assessment practice at School A.....	150
5.7.2 Tentative generalisations on the relationship between formative and summative assessment.....	151

Chapter 6 Case study B: Changes to the relationship between formative and summative assessment

6.1 Introduction	153
6.1.1 Chapter overview	153

6.1.2 School B context	154
6.1.3 School B data and analysis	156
6.2 Whole school processes (W).....	159
6.2.1 DBR Phase 1W - Formative or summative purpose.....	159
6.2.2 DBR Phase 2W - Formative for summative purpose	162
6.2.3 DBR Phase 3W - Formative and summative purpose: assessment as ongoing.....	163
6.2.4 Summary of changes at whole school processes level	165
6.3 Summative reporting layer (S).....	166
6.3.1 DBR Phase 1S and 2S - A focus on records and evidence	166
6.3.2 DBR Phase 3S - Confidence in teacher judgement	169
6.3.3 Summary of changes at summative reporting layer.....	172
6.4 Monitoring layer (M).....	172
6.4.1 DBR Phase 1M - Concern for consistency	172
6.4.2 DBR Phase 2M - Levelling and moderation	173
6.4.3 DBR Phase 3M - Range of information	175
6.4.4 Summary of changes at monitoring layer.....	177
6.5 Teacher layer (T)	177
6.5.1 DBR Phase 1T – Strategies include marking	177
6.5.2 DBR Phase 2T - Making assessment manageable.....	179
6.5.3 DBR Phase 3T - More open	183
6.5.4 Summary of changes at teacher layer	184
6.6 Pupil layer (P).....	185
6.6.1 DBR Phase 1P - Trialling strategies	185
6.6.2 DBR Phase 2P – Developing strategies for self and peer assessment	186
6.6.3 DBR Phases 3P - Developing the role of the pupil	188
6.6.4 Summary of changes at pupil layer.....	190
6.7 Conclusion	191
6.7.1 Key features of changing assessment practice at School B	191
6.7.2 Tentative generalisations.....	192

Chapter 7 Discussion

7.1 Introduction.....	193
------------------------------	------------

7.2 Summary of key findings in response to research questions	194
7.2.1 Formative and summative assessment in primary science (RQ1)	194
7.2.2 Relationship between formative and summative assessment (RQ2).....	195
7.2.3 Change over time (RQ3)	195
7.3 Conceptualisation and enactment of the relationship between formative and summative assessment.....	196
7.3.1 Dimensions in the conceptualisation and enactment of the relationship between formative and summative assessment	196
7.3.2 A distinction to support the relationship between formative and summative assessment: summative assessment as snapshot or summary	203
7.3.3 Strategies for teacher assessment	204
7.3.4 Summary of conceptualisation and enactment of the relationship between formative and summative assessment	206
7.4 Validity and reliability in ‘formative to summative’ assessment	207
7.4.1 Discussion of validity and reliability in ‘formative to summative’ assessment .	207
7.4.2 A seesaw model of teacher assessment.....	211
7.4.3 Summary of validity and reliability in ‘formative to summative’ assessment...	215
7.5 Representation of the relationship between formative and summative assessment in the TAPS pyramid	216
7.5.2 Summary of the representation of the relationship between formative and summative assessment in the TAPS pyramid	220
7.6 Summary	221

Chapter 8 Conclusions and recommendations

8.1 Introduction	222
8.2 Key findings and reflections on the study	223
8.2.1 Key findings and products.....	223
8.2.2 Reflections on the advantages and limitations of the DBR process.....	226
8.2.3 Summary of reflections on advantages and limitations of DBR process.....	229
8.3 Recommendations	230
8.3.1 Recommendations for practice	230
8.3.2 Recommendations for policy	232
8.3.3 Recommendations for research	233

Appendix

References	235
Chapter 3 Appendices	248
3A: Collection of Assessment Samples	248
3B: TAPS cluster day 1 formative/summative written task Oct13	248
3C: Impact Questionnaire to TAPS project teachers June 2015	249
3D: Lesson observation schedule/framework, School visit 1 Nov13	249
3E: Lesson observation schedule/framework, School visit 3, March 2014	251
3F: Interview with science subject leader, school visit 1, November 2013	252
3G: Interview questions for science subject leader, June 2016.....	253
3H: TAPS project consent forms	253
Chapter 4 Appendices	254
4A: PSQM Round 4 data	254
4B: PSQM initial coding.....	255
4C: PSQM second level of analysis for summative assessment	2575
4D: PSQM second level of analysis for formative assessment	256
Chapter 5 Appendices	258
5A: School A case record.....	258
5B: School A case study codes	261
5C: School A case study codes organised by TAPS pyramid layers.....	262
Chapter 6 Appendices	263
6A: School B case record.....	263
6B: School B case study codes	266
6C: Themes.....	267
6D: Card Sort.....	268
6E: List of conferences/teacher presentations (Jan 14 – July 16).....	269

List of figures

Figure 1.1 Mapping of key data and outputs from PhD and TAPS	25
Figure 2.1 The Nuffield data-flow pyramid model (Nuffield 2012: 20)	42
Figure 2.2 TAPS pyramid school self-evaluation tool (Davies et al. 2014)	44
Figure 2.3 TAPS pyramid school self-evaluation tool (Earle et al. 2015)	45
Figure 3.1 TAPS pyramid analytical framework: pyramid layers and themes	90
Figure 4.1: Summative assessment (detailed) methods for PSQM Round 4	96
Figure 4.2: Summative assessment (summary) for PSQM Round 4	96
Figure 4.3: Elicitation strategies mentioned in reflections for PSQM Round 4	104
Figure 4.4: How self-assessment was described by the 33 schools	108
Figure 5.1 TAPS pyramid analytical framework	120
Figure 5.2 Science Skills Wheel	129
Figure 5.3 School structures	132
Figure 6.1 TAPS pyramid analytical framework	157
Figure 6.2 Strategies sorting card activity	160
Figure 6.3 Pyramid extract, Feb 2014	167
Figure 6.4 Marking example	177
Figure 6.5 End of lesson expectation grid	180
Figure 6.6 Self-evaluation sticker at end of topic on water cycle	185
Figure 6.7 Key Stage 1 science toolkit	186
Figure 7.1 The Teacher Assessment Seesaw	210
Figure 7.2 Teacher Assessment Seesaw for summative tests in primary science	212
Figure 7.3 Teacher Assessment Seesaw for summative tasks for all objectives	213
Figure 7.4 The ‘formative to summative’ process in the TAPS pyramid model (detailed)....	216
Figure 7.5 The ‘formative to summative’ process in the TAPS pyramid model (summary)..	218
Figure 8.1 PhD summary	222

List of tables

Table 2.1 Summary of issues and areas for focus.....	56
Table 3.1 Design-Based Research Phases mapped onto research questions and key data...	63
Table 3.2 Overview of data collection methods.....	69
Table 3.3 Theory-led and emergent codes.....	86
Table 3.4 Overview of qualitative data analysis.....	87
Table 5.1 Lesson observation summary.....	141
Table 6.1 Design-Based Research Phases.....	156
Table 6.2 Formative and summative written task, TAPS cluster day 1 October 2013	158
Table 7.1 Dimensions in the teachers' conceptualisation and enactment of assessment....	196
Table 7.2 Dimensions in the relationship between formative and summative assessment.	200
Table 7.3 Examples of different types of teacher assessment strategies.....	204
Table 8.1 Key findings and products.....	224

Chapter 1 Introduction

1.1 Introduction to key concepts

Assessment is a powerful driver in education: it influences school and classroom culture; it determines what is taught and how; and it directly impacts on pupil and teacher conceptualisations of learning (Edwards 2013). Assessment is a complex, embedded and integral part of teaching with a multitude of strategies, purposes and consequences. Pupil experience of primary science is shaped by assessment practices, thus it is essential for such practices to be well understood. This study analyses the conceptualisation and enactment of science assessment within primary schools.

This chapter will set the scene for the research by defining key concepts and explaining the rationale and context for this thesis. Key projects and organisations relevant for the research will also be introduced. In this section, key terminology will be briefly defined to support the following discussion, with all terms being examined more closely in Chapter 2.

Assessment is the part of teaching where a judgement is made regarding learning. It includes the process of collecting and interpreting evidence to make judgements about pupil achievement (Harlen 2007: 11). Edwards (2013) describes assessment as: *'an integral part of learning'* and: *'a key component in quality teaching'* (p213). It often includes a judgement against a reference point, which might be a previous performance (ipsative), a peer (norm-referenced) or a set of criteria like the National Curriculum (criterion-referenced) (Gipps 1994), with more recent assessment initiatives favouring the latter (Lum 2015).

The purpose of **formative assessment** is to inform decisions about learning experiences (Harlen 2007), to seek out and respond to information to enhance ongoing learning (Klenowski 2009). It is the part of classroom practice where the teacher or pupil checks their progress and considers what to do next; it is a process which promotes pupil learning (Harrison and Howard 2009). Strategies associated with formative assessment include:

identifying and making explicit success criteria; elicitation of children's existing ideas; feedback; self- assessment and peer assessment (Wiliam 2011). In order to emphasise the requirement to have an impact on learning, 'Assessment for Learning (AfL)' became a popular way to refer to formative assessment (Assessment Reform Group 1999), which will be discussed further in Chapter 2.

The purpose of **summative assessment** is to report or summarise attainment, for example, a report to parents or an end of topic activity which is designed to encapsulate what the pupil has learnt in relation to the topic or 'learning objectives'. In some cases this information is used to hold schools to account (Whetton 2009), with assessment results becoming 'high stakes' when they are used for school target setting and the ranking of schools, a proxy for judgements of the education system (Mansell et al. 2009). Taras (2005) argues that all assessment begins with a summative judgement and the distinguishing characteristic between formative and summative is whether there is feedback which is acted upon. The relationship between formative and summative assessment is the key focus for this study.

Three key concepts for the evaluation of assessment are introduced in turn below: validity, reliability and manageability.

Validity concerns whether an assessment is actually assessing what it claims to and the extent to which it is fit for purpose (Green and Oates 2009), for example, whether an assessment of primary science effectively samples enough of the domain to be representative (Stobart 2009). Validity is primarily about purposes, for example, if the purpose of formative assessment is to stimulate further learning, then the assessment will only be valid if further learning is supported (Stobart 2012). Validity has been described as the most important consideration for assessment procedures (Crooks et al. 1996) and includes a number of facets which will be considered further in Chapter 2.

Reliability concerns trust in the consistency of an assessment (Mansell et al. 2009), for example, whether the same result would be found if the task was given on a different occasion or marked by a different teacher (Newton 2009). Inter-rater reliability is often the focus for discussions of this strand (Black and Wiliam 2012), but Johnson (2013) also notes

that lack of clarity and applicability of assessment criteria leads to unreliability. In order to be valid, an assessment needs to reliably assess what it has been designed to, so reliability is a necessary condition of validity, but it is not sufficient, since to be valid an assessment also needs to sample enough of the domain. The relationship between reliability and validity in the teacher assessment of primary science is an important area (Harlen 2013) and a focus for consideration in this study.

Manageability is a key principle underpinning 'quality assessment' (Edwards 2013) because the assessment practices need to be perceived as manageable by teachers if they are to be enacted. For example, practices which are deemed to require too much teacher time are likely to be dropped in favour of things which will be easier to implement, thus manageability was found to be a key concern for teachers (Davies et al. 2014).

The key terminology defined briefly above will be considered in greater depth in Chapter 2. The next section will consider why this research is important and what it aims to achieve.

1.2 Rationale and aim

1.2.1 Rationale

Assessment is fundamental to the practice of education, yet it is not neutral, it is value-laden; assessment processes determine what is 'valuable to learn' and what success will look like, they: *"creates and shapes what is measured"* (Stobart 2008: 1). Since assessment shapes the curriculum as experienced by children, it is essential for such assessment practices to be well understood by teachers. However, assessment has been identified as the weakest aspect of teacher practice (Black and Harrison 2010). Assessment can encourage learning, or it can undermine it (Stobart 2008). Researchers point to the harmful effects of poor assessments or misinterpretation of assessment purposes (Mansell et al. 2009, Murphy et al. 2013, Boaler 2015, Black 2012).

The functions and effect of assessment have received much attention, with some arguing (e.g. Black and Wiliam 1998) that assessment should have an impact on learning otherwise

there is little point in conducting the assessment in the first place. Research into formative assessment champions the use of assessment to support learners with their next steps (Gardner et al. 2010); whilst summative assessment became viewed in a negative light because of suggestions that it was the cause of curriculum narrowing and teaching to the test (Harlen 2013). However, education systems require both purposes to be fulfilled, with assessment information used to both support learning and to summarise achievements for a range of audiences such as pupils, parents, senior leaders and the next class teacher. Such a clash between a positive view of formative assessment and a negative view of summative assessment may be counter-productive, leading teachers to run separate, and consequently unmanageable, assessment systems (Earle 2014). The enacted relationship between formative and summative assessment has implications for both the validity and reliability of teacher assessment, with Johnson (2013) noting a lack of evidence and research in this area.

Assessment has increasingly become a political issue, with international comparison of student achievement data leading governments to implement assessment reforms and standards-driven curricula (Connelly et al. 2012: 593). A major ongoing concern in England, which will be discussed further in Section 1.3, is the lack of centralised guidance for primary teachers on how to make valid and reliable teacher assessment judgements of primary science (Turner et al. 2013: 3). 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), and consequently 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 the relationship between formative and summative assessment.

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 practical activities. Teacher judgement can take into account a range of outcomes which are not easily assessed in a test. 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 other teachers. Wiliam (2003) describes an inevitable 'trade off' between reliability and validity, and argues that teacher assessment can be made more reliable. Harlen (2007) proposes that 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, although this raises issues of manageability.

A closer relationship between formative and summative assessment is seen by some as crucial to effective teacher assessment (Wiliam and Black 1996, Hodgson and Pyle 2010, Nuffield Foundation 2012, Harlen 2013), thus a particular focus for this study will be to explore conceptualisation and enactment of this relationship within schools.

1.2.2 Aim and research questions

Aim

The aim of this research is to develop understanding of the relationship between formative and summative assessment in action, in order to inform guidance for practice to support teacher assessment in primary science.

This aim situates the research within an 'Integrated Knowledge Tradition' (Furlong and Whitty 2017), whereby academic and practical knowledge are brought together, considering 'knowing how' in practice as well as 'knowing that' theoretically. Such research aims to improve practice through engagement in real world settings, utilising empirical enquiry in cycles of development over time (Furlong and Whitty 2017: 43). Methodological consideration of such applied research will be the main topic of discussion in Chapter 3.

In order to fulfil the aim described above, three research questions (RQ) are proposed; these will be considered more closely in Chapter 3, but are presented here to support the introduction to the research.

Research questions

RQ1. How do teachers assess children’s learning in science for **formative and summative purposes**?

RQ2. How can teachers’ conceptualisation and enactment of the **relationship between formative and summative** assessment of children’s learning in science be used to inform guidance for practice?

RQ3. How can study of **changes over time** in conceptualisation and enactment of the relationship between formative and summative assessment be used to inform guidance for practice?

1.3 Context

1.3.1 Assessment in England

Primary teachers in England have a statutory requirement to summatively assess each child against the National Curriculum descriptors in English, maths and science at ages 7 and 11 (DfE 2013a, STA 2015). 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 maths (see Section 1.3.2).

During the period of this study, 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; together with 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 is now 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; such a change has had significant implications for this study, which explored assessment practice during this time of change.

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' and topics of biology, chemistry and physics such as: plants, everyday materials and electricity. Guidance explicitly states 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'. The Commission on Assessment without Levels (2015) noted however 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'. It may take some time for valid, reliable and manageable systems of teacher assessment of primary science to emerge since this change took place within a political environment that eschewed prescription and espoused schools freedom to develop their own responses.

The new assessment arrangements may take several years to become an established feature of classroom practice. Time has been noted by many 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) noted, before the recent removal of levels, 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 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 maths, since science currently does not feature in accountability measures like the ‘floor standard’ which could be the trigger for an Ofsted inspection. If science assessments are not high stakes, then it follows that there should be less 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.

1.3.2 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 maths (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 maths (e.g. Ofsted 2013, Wellcome Trust 2014, CBI 2015).

Eady's (2008) study of the purpose of teaching science in primary school found that many teachers saw scientific knowledge entwined with 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. 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 which 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 they are teaching primary science.

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 CBI (2015). Whilst 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 for scientific enquiry which: '*sustains pupils' natural curiosity*' (Ofsted 2013: 5). In terms of this research, the current climate, with its limited time for the teaching of primary science, means that manageability of assessment processes will need to be a key priority for any recommendations.

Concerns have also been raised about the support 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 are able to devote to this.

1.3.3 My own professional context and experience

My own professional experience also provides a context and rationale for this research. As a primary school teacher (1999-2012), I was both science subject leader and assessment coordinator and spent a lot of time supporting colleagues with teacher assessment, although the statutory requirements of end of Key Stage assessment dominated. More recently as a teacher trainer, working with both experienced and trainee teachers (2007-present), assessment has been a concern for all, with constant changes to centralised guidance and requests for further support.

Professional development courses, which I have attended or delivered, have focused on the development of strategies for formative assessment and statutory requirements for summative assessment, without exploration of the relationship between the two. Thus my professional experience also indicates a need to research the relationship between formative and summative assessment.

1.4 Links between PhD and key projects

1.4.1 Key projects and organisations

This study is strengthened by its involvement with one key project and two organisations which will be introduced briefly below.

The **Teacher Assessment in Primary Science (TAPS) project** is an ongoing research project based at Bath Spa University and funded by the Primary Science Teaching Trust. It began in September 2013 with the remit to address the concerns raised in Section 1.3, that there was a lack of support for assessment in primary science. TAPS operationalised a model of science assessment put forward by the Nuffield Foundation (2012), creating a pyramid-shaped school self-evaluation tool (Davies et al. 2014) which proposed utilising formative teacher assessment for summative purposes, to support valid, reliable and manageable summative judgements. TAPS is not the only assessment research relevant to this study; the field will be reviewed in more detail in the next chapter, but it is significant here because it provides the wider project background to this study. The relationship between TAPS and the PhD will be considered more closely in the next section.

The **Primary Science Teaching Trust (PSTT)**, as well as funding the TAPS project, also supports teachers via clusters, academic collaboration projects and the creation of College Fellows through the presentation of Primary Science Teacher Awards each year. The PSTT College Fellows have been an additional source of data for this study in terms of triangulation and ongoing feedback on the development of guidance and resources.

The **Primary Science Quality Mark (PSQM)** is an award scheme to enable primary schools to develop science leadership, teaching and learning (White et al. 2016). It requires the science subject leader 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 PSQM database was one of the sources of data for this study.

1.4.2 Relationship between TAPS and PhD

This PhD study overlaps and complements the larger TAPS research project in terms of time, data and personnel, each of which will be discussed further below. The key datasets and outputs will be explained more fully in the ensuing chapters, but are included here to clarify the relationship between my study and TAPS.

This PhD research began in January 2013 with analysis of the PSQM database to map approaches to assessment. In September 2013 this data analysis was one of the sources of evidence used to inform the initial stages of the TAPS project. Thus the PhD and TAPS have been linked from the outset, and the PhD research continued to be a major part of the TAPS project.

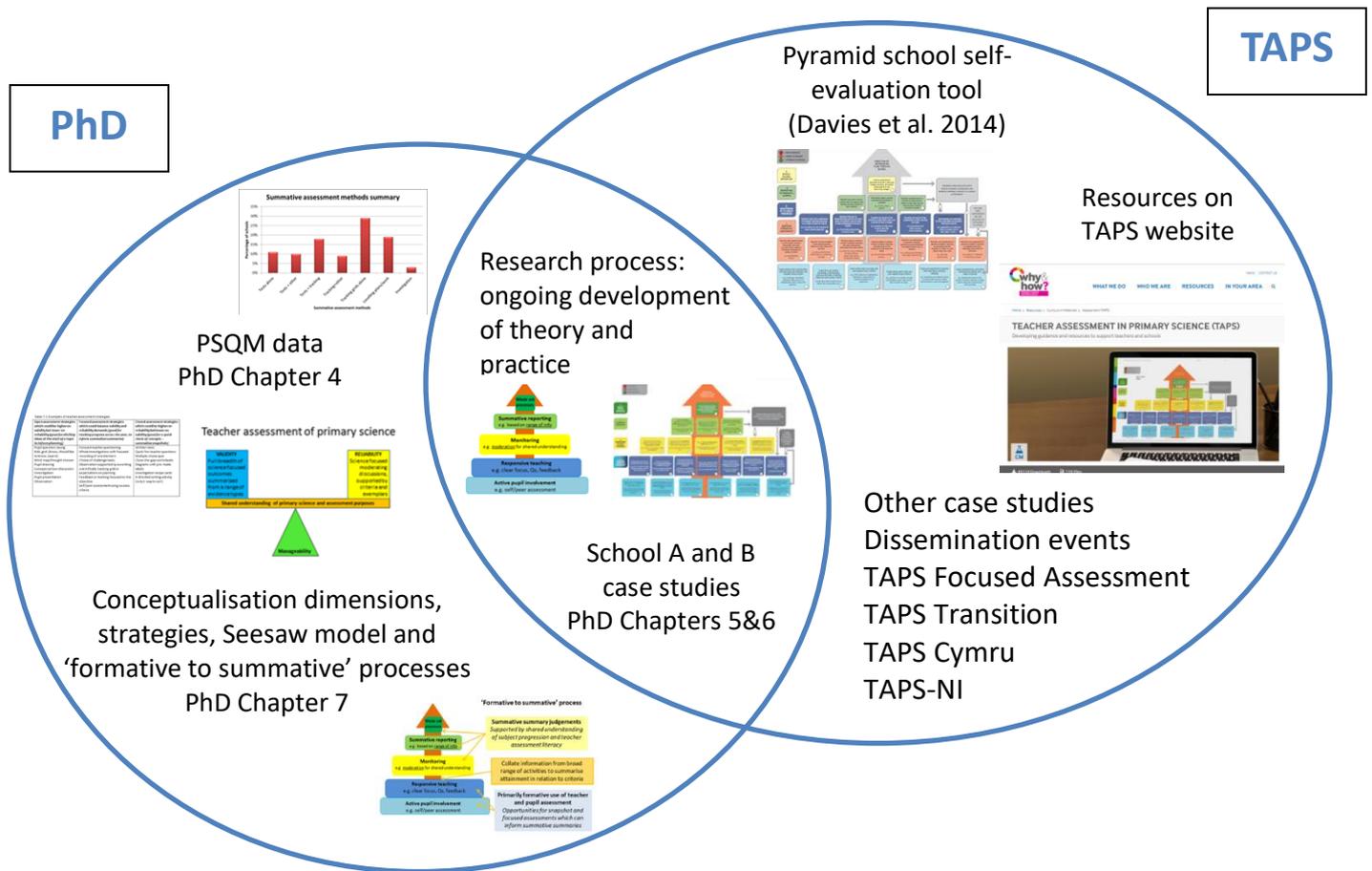
The initial TAPS project (TAPS1: 2013-16) involved twelve project schools working with a research team of 5 tutors; I was first part of the research team and then became the project lead from January 2015. The PhD provided a greater depth of study and TAPS provided a wide range of critical friends (both researchers and teachers) and complementary data, to support and extend the study. The work presented here is original since it arises from my own data collection and journey through the project.

This PhD study utilises some data from the TAPS project, for example, I was link tutor for TAPS project schools, collecting data at TAPS cluster days and on TAPS school visits. The schools' PSQM submissions also became part of the PhD data, in addition to TAPS. Such dual purposes were discussed with the teachers, to ensure ethical processes were followed. The TAPS data was analysed afresh for the PhD, separately to the TAPS project, with a more in-depth focus on the relationship between formative and summative assessment for this PhD study.

A key outcome of the TAPS project was the development of a pyramid-shaped model of school self-evaluation, with layers from classroom practice to whole school processes (Davies et al. 2014). In this study the TAPS pyramid layers are used as an analytical framework to provide a comprehensive analysis of assessment practice in the case study schools. This PhD study also provided the opportunity to critically analyse the use of the TAPS pyramid with the aim of refining the model and developing further guidance.

Figure 1.1 contains a mapping of some key data and outputs from both the PhD and TAPS to support the reader to visualise the way the PhD and TAPS have overlapped and supported each other.

Figure 1.1 Mapping of key data and outputs from PhD and TAPS



The PhD and TAPS are overlapping and complementary, with each informing the other. The aim of this PhD is to focus closely on the relationship between formative and summative assessment to develop guidance for practice; such recommendations could be utilised in the ongoing development of the TAPS project.

1.5 Overview of thesis

This chapter set the scene for the research by first defining the key concepts of: validity, reliability, manageability, formative and summative assessment. The context of primary science assessment in England was introduced, with the problematic relationship between formative and summative assessment, and the low status of science in primary schools. The

relationship between the PhD and the TAPS project was explained and summarised in Figure 1.1. The rest of the thesis can be summarised as follows:

- Chapter 2 will review literature to explore the relationship between formative and summative assessment.
- Chapter 3 will consider methodology appropriate to bring academic and practical knowledge together.
- Chapter 4 will map current practice in formative and summative assessment of primary science (RQ1) through analysis of the Primary Science Quality Mark database.
- Chapter 5 will focus on conceptualisation and enactment of the relationship between formative and summative assessment in primary science (RQ2) through the development of a case study of one school.
- Chapter 6 will consider changes over time in conceptualisation and enactment of the relationship between formative and summative assessment (RQ3) through the development of a case study of a second school.
- Chapter 7 will discuss how the findings can be used to develop guidance for practice (RQ2 & RQ3).
- Finally, Chapter 8 will summarise recommendations for practice, policy and research.

Chapter 2 Literature Review

2.1 Introduction

Assessment is both ‘an integral part of the educational process’ (DES 1988) and a powerful driver which can impact all areas of schooling. Assessment has an impact on the culture within the school since it affects what is understood by learning and what is worth learning (Edwards 2013: 213). It influences both the teaching and those being taught: *“Assessment does not objectively measure what is already there, but rather creates and shapes what is measured”* (Stobart 2008: p1). The power of assessment is recognised, but there is not agreement about its implementation. On the one hand, assessment has been identified as: *“one of the most powerful tools for promoting effective learning”* (ARG 1999: 2), whilst at the same time the list of negative consequences of assessment has been growing, for example: narrowing of the curriculum; the labelling of children; an increased focus on accountability at the cost of pupil learning and wellbeing (Wiliam 2003, Boaler 2015, Whetton 2009, Broadfoot 2007, Murphy et al. 2013).

The competing uses of assessment places the teacher in a ‘conflicted position’, with assessment for accountability seeming to require a different approach to using assessment as part of the learning process (Green and Oates 2009: 233). Lum (2015) suggests that a recent paradigm shift, from assessment for the purpose of comparison, to assessment to support learning, has changed the way assessment is perceived. This chapter will explore assessment theory in relation to teacher assessment in the current context, in particular considering ‘reliability’, which concerns the extent to which an assessment can be trusted to give consistent information, and ‘validity’, which concerns whether the assessment measures: *‘all that it might be felt important to measure’* (Mansell et al. 2009: 12). The balance of these two concepts is a key area for debate regarding formative and summative purposes in teacher assessment.

This chapter will begin with an exploration of validity and reliability in teacher assessment (Section 2.2). This will be followed by analysis of the distinctions made between formative

and summative assessment (Section 2.3), before discussing a ‘formative to summative’ approach, which has been put forward to balance the demands of validity and reliability in teacher assessment. The final literature section (2.4) will explore the nature of primary science education, since before deciding whether an assessment is valid, it is necessary to understand the domain of the assessment. The chapter will end (Section 2.5) by summarising key issues which have arisen from the literature review, feeding into the presentation of three research questions.

2.2 Validity and reliability in teacher assessment

2.2.1 Validity

Validity concerns whether an assessment does assess what it claims to (Green and Oates 2009), however, it is a complex and multi-faceted concept. Validity is not the ‘static property’ of an assessment, which is either there or not, it is contingent on the purpose(s), use(s) and interpretation(s) of the assessment (Stobart 2009). In order to explore the validity of teacher assessment in primary science, a number of features of validity are explored in turn below.

Both content and construct validity insist that the assessment measures what it is meant to, the conceptual content and the skills for a particular subject. **Content validity** concerns how well the agreed curriculum is sampled, whilst **construct validity** concerns how well this represents the underlying skill or concept (Stobart 2009). For example, in the case of primary science, a multiple choice test on the digestive system would not validly assess the full range of inquiry skills and understanding of the whole of practical primary science; however, it could be combined with other assessments to provide a fuller picture of pupil performance. Construct validity goes beyond content validity because it may challenge the way the curriculum represents a construct. For this study, this means that it is necessary to explore the nature of primary science (Section 2.4) as well as how it is assessed.

An important consideration is the ‘sampling’ of the construct, acknowledging that the assessment will sample the construct in a particular context or particular format, it is not measuring the construct directly; assessment is an approximation. Nevertheless, what is

sampled can affect both the validity of the assessment and the teaching of the subject. Messick (1989) suggests that the best protection against invalid assessment inferences is to minimise construct under-representation and construct-irrelevant variance. **Construct under-representation** is a threat to validity (Black and Wiliam 2012), especially in primary science, since the key inquiry skills of Working Scientifically are arguably much harder to assess, being less accessible via a written test for example, thus they are likely to be under-represented in classroom assessments. Maintaining a focus on the science is important to avoid **construct-irrelevance**, for example, a teacher marking pupil work may comment on the neatness of handwriting or use of grammar, thus assessing writing skills rather than science inquiry skills. Stobart (2009) also notes that predictability in assessment can be a form of construct irrelevance, since it could be an assessment of 'rote learned responses' rather than a 'demonstration of skills' (p168).

Predictive validity requires that the assessment provides accurate predictions for the outcomes of future assessments (Isaacs et al. 2013: 137). This has arguably become more important in primary schools as teachers have increasingly been asked to predict the performance of children at the end of the year or Key Stage on the basis of 'mock' tests. Under a numerical or levelling system this prediction or target could be worked out via a formula, for example, moving up two-thirds of a level per year. Such use of assessments appears to rely on dubious assumptions: that learning is linear with children moving in a predictable way along the continuum each year (Stobart 2008). Alternatively, the assumption could be made that the assessment is measuring some underlying 'ability' within the child which will remain stable as they move through school; such innate fixed intelligence remains a topic of fierce debate in education circles, which unfortunately there is not space to discuss here. Suffice to say that Stobart (2009) warns of over-simplistic interpretations of assessments which treat the result as a direct measure of an underlying educational standard (p175). Measures of predictive validity will not be a focus for this study, because that would require a focus on pupils over time, and the concern for this study is for teacher practice over time.

Gardner et al. (2010) assert that: "*assessment of any kind should ultimately improve learning*" (p2), which suggests that the impact of assessments should be judged by the way

which they impact learning. Information gathered for formative purposes will have **consequential validity** if it is used to support the learning of the pupils. This is the primary aim of Assessment for Learning, but it is also possible for there to be pre-emptive assessment (Carless 2007, cited in Stobart 2012: 233), whereby it is future pupils who are benefitting since it is the teacher who learns from the assessment and considers how to adjust their teaching for the next cohort. However, it is hard to imagine that every summative assessment can be said to have a direct, or even indirect, impact on learning. Thus it could be argued that consequential validity applies more to formative rather than summative assessment, or that consequential validity is a measure of how successful the assessment was in meeting its purpose. Unintended consequences of an assessment should also be considered, for example, an end of Key Stage test may meet its purpose of providing a summary level for parents and schools, but the social consequences of such testing could include a narrowing of the curriculum and ‘teaching to the test’ which leads to a reduction in the uptake of the subject in secondary school. It is interesting to consider how responsible test constructors are for the use of their tests; it feels unrealistic to expect that test validity only be decided once the cohort taking the test has completed their schooling. Stobart (2009) suggests that: *“any validity argument begins and ends with purposes”* (p166); with validity so bound up with assessment purposes, it is necessary to return to this discussion below when considering formative and summative assessment in more detail.

2.2.2 Reliability

Reliability concerns trust in the accuracy or consistency of an assessment (Mansell et al. 2009). This can be internal, within the assessment, or external, between assessments. Although Filer and Pollard (2000) caution that since school assessment necessarily takes place in a social context, the presumed ‘objectivity’ of some assessments is actually a myth: no assessment can be perfectly objective, repeatable and reliable.

Internal issues with tests or tasks such as the wording of questions and the conditions under which they are taken can be sources of unreliability (Johnson 2012: 68), but the use of a range of information in teacher assessment aims to mitigate issues with particular tasks. Consistency between test items and test conditions are perhaps less of a concern with the

move away from pencil and paper tests (Green and Oates 2009), towards lists of competencies or tasks. However, where stand-alone tasks are performing a similar snapshot function, of recording attainment at a particular point in time, then the construction of such tasks should be examined.

Standish (2007) suggests that the 'vogue' for lists of competencies is linked to a new behaviourist revival, with observables regaining focus for teaching and assessment. A focus on behaviour could lead to an exaggerated focus on performance, with a: *"denial that anything is learned unless it can be demonstrated in clearly measurable learning outcomes, and these are clearly specified in advance"* (Standish 2007: 168). Torrance (2005, cited in Stobart 2008: 157) points out that '**criteria compliance**' can follow when objectives are too detailed, leading to 'assessment as learning' where the goal becomes surface level ticking or highlighting of a large number of criteria rather than in-depth learning: a 'tick-box culture' (Mansell et al 2009). This also links to the mistaken assumption that frequent summative testing will support learning: *"Marks, levels, judgmental comments or the setting of targets, cannot, on their own, be formative. Pupils may need help to know how they can improve"* (Mansell et al 2009: 10). Assessment judgements are necessarily based on outward behaviours, but the way such 'lists of competencies' are utilised in practice will need to be explored in this study.

A focus for reliability in UK teacher assessment is on **inter-rater reliability** (Black and Wiliam 2012: 247), which addresses whether the same judgement would be made by different teachers on the basis of the same set of evidence. Johnson (2013) asserts that there is a lack of evidence on the reliability of teacher assessment, although: *"potentially the most effective strategy for ensuring both validity and reliability in teacher assessment, if these can in principle be achieved to an acceptable degree, is consensus moderation"* (p99).

Moderation will be explored further below, suffice to say that in primary science inter-rater reliability would perhaps be enhanced by developing a shared understanding through moderation and exemplar material. Assessment with the sole purpose of formatively supporting the pupil with their next steps would arguably be less concerned with reliability (Harlen 2007), since comparison with others is not the prime purpose and the pupil is likely to have moved on in their learning before another 'rater' attempts to assess their learning.

However, without an idea of progression in scientific skills and understanding, then the teacher may find it difficult to support the child to progress. *“Moderated teacher assessment has been proven to facilitate staff development and effective pedagogic practice”* (Green and Oates 2009: 238). It appears that there needs to be some common understanding of what it looks like to ‘be better’ at science to be able to fulfil both formative and summative purposes of assessment.

Assessment requires a judgement against a reference point, that reference point might be a previous performance (ipsative), a peer (norm-referenced) or a set of criteria like the National Curriculum (criterion-referenced) (Gipps 1994). However, Halliday (2010) suggests that these definitions may not be distinct, for example, when norm-referencing the teacher will also be drawing upon some sense of criteria, and when criterion-referencing, it is useful to have a sense of what an ‘average’ performance would look like. Assessment reference points are contained within a social context. There may also be more than one reference point, with multi-criterion comparisons informing teacher judgements and feedback (Sadler 1989). Davis (1998) suggests that criterion-referenced assessment requires a ‘shared conception of achievement’, something that some would argue needs to be built within a community. For this study, it will be useful to consider what types of referencing the teachers are using for their judgements.

2.2.3 Validity and reliability in teacher assessment

Teacher assessment is the term used to describe assessment practice whereby the teacher makes the judgement regarding pupil attainment; this may be on the basis of one task or, more commonly, a range of tasks and evidence. Mansell et al. (2009) and Gardner et al. (2010) argue that teacher assessment is a more valid means of summative assessment than testing because it can reduce construct under-representation, providing a broader sampling of the construct by taking into account the wide range of information available in the classroom. Teacher judgement can take into account a range of learning processes and outcomes which are not easily assessed in a test; this is particularly important for science since its essence is practical, scientific inquiries can utilise dialogue, collaboration, practical skills and problem solving. However, the large-scale collection of evidence to support such teacher assessment (Harlen 2007), prompts questions of manageability for teachers.

Torrance and Prior (1998) describe convergent and divergent teacher assessment. Convergent assessment aims to discover whether a pupil knows a predetermined thing, whilst divergent assessment aims to discover what the learner knows or can do. The classroom teacher may relate such continua to open or closed tasks, with divergent open tasks providing broader assessment information, with possibly higher validity, whilst closed convergent tasks are more tightly focused allowing them to be more reliably judged. There are also 'windfall' opportunities where the pupils choose to use a particular skill without particular probing from the teacher (Black and Wiliam 1996). This raises questions for this research, in terms of the particular types of assessment which teachers should utilise and how divergent or convergent such teacher assessment should be.

The role of the teacher includes: structuring the situation, modelling, instructing, giving feedback, questioning and structuring concepts (Tharp and Gallimore 1988: 34, 47). The amount of structure and support is a key concern in the field of teacher assessment: should the judgement be based on aided or unaided attainment. The amount of support has implications in terms of reliability, for consistency of judgements, and in terms of validity, with regards to whether the assessment is assessing what it purports to.

From a social-constructivist perspective, Vygotsky described learning in terms of a Zone of Proximal Development (ZPD), where a pupil moves from a level of actual development to their level of proximal or potential development (Alexander 2008). Feedback is the information about the gap between the actual level and the reference level of the system parameter, which is used to alter the gap in some way (Ramaprasad 1983: 4 cited in Taras 2005). Black and Wiliam (1996) note the importance of using the information, it is only considered as feedback when it is used to alter the gap. For such information to be useful in closing the gap, there needs to be a sense of what must be changed, a relation to a 'developmental model of growth', effectively construct-referenced (Messick 1975, cited in Black and Wiliam 1996). Vygotsky (1978) described the actual developmental level as the end product since the functions have already matured, the 'fruits of development' (p86) and it is this independent performance which is assessed via testing. However, Vygotsky argued that focusing on the 'actual level' is retrospective since development has already been

completed, there is more predictive power in considering the ‘buds or flowers of development’ (1978: 86). This is perhaps particularly pertinent when considering assessment in primary science where most practical work is carried out in groups, which may include a ‘more knowledgeable other’ leading the way. This discussion is echoed in Gipps (1994) who explores whether assessment should take account of best or typical performance (p9). The issue of what is to be assessed and whether this is independent ‘actual development’ or collaborative action in the ZPD will need to be considered in the course of this research.

Teacher assessment may provide the opportunity for increased validity, but: *“teachers’ assessments are often perceived as having low reliability”* (Harlen 2007:25, Black et al. 2011). Johnson (2013) questions reliability in teacher assessment, noting the limited and ambiguous research in this area. One concern is the permanence of assessment evidence: whether it is recorded in some way, which can then be considered for consistency by other ‘raters’; or whether it is, for example, group discussion which is harder to capture. Black and Wiliam (1996) argue that inter-rater consistency is not important for formative assessment, and that both written and oral accounts are ‘imperfect representations’ of the pupil’s thoughts. Connelly et al. (2012) discuss how teacher judgements draw on multiple sources of knowledge and evidence, so they are doing much more than a matching of evidence to criteria. Such judgements must be considered in context, taking into account teacher beliefs, attitudes and practices; for example, teachers will draw on their tacit knowledge of students and previous evaluative experiences. However, such an ‘expansive’ model of teacher assessment (Lum 2015), where teachers make judgements based on a wide range of evidence, could be open to concerns regarding subjectivity and teacher bias (Campbell 2015).

Moderation discussions, where teachers compare and analyse judgements, provide the opportunity to support teachers to make their tacit knowledge explicit (Sharpe 2004). Klenowski and Wyatt-Smith (2014) suggest that enhancing consistency of judgements is only one half of the purpose of moderation; a second goal is to improve the teachers’ assessment and pedagogical practice: their assessment competence or literacy (Black et al. 2011). Through social moderation, involving discussion and debate of evidence of pupil

outcomes: *“a shared understanding of the standards is negotiated”* (Klenowski and Wyatt-Smith 2014: 75). Black et al. (2011: 458) found that teachers could learn to use more holistic judgements rather than rely on a prescriptive tick list. Connelly et al. (2012) found that the majority of teachers involved in their study welcomed the peer support and believed that the moderating discussions resulted in consistency of teacher judgement; however, some teachers described more negative responses regarding the time and effort involved in such a complex process.

2.2.4 Relationship between validity and reliability in teacher assessment

William (2003) argues that there is inevitably a ‘trade off’ between reliability and validity. Halliday (2010: 370) asserts that a ‘trade off’ between reliability and validity is necessary since reliability relies on a narrowing of task variables to support marker agreement, whilst validity depends on the opposite: as broad a sampling of the subject as possible. Sadler (1989) asserts that validity should take precedence when the aim is formative, for diagnosis and improvement (p122). Davis (1998: 140) suggests that high reliability and validity are possible, but only if a ‘very narrow kind of achievement’ is examined. However, Stobart (2009: 168) describes reliability as an ‘essential part’ of validity, rather than a separate component, since poor reliability threatens validity. Nevertheless, he goes on to argue that a search for ‘maximum reliability’ may limit what can be measured, thus reducing construct validity. So it would appear that for an assessment to be valid, it requires a certain amount of reliability, but a focus on only the latter is likely to reduce the validity overall: there is a ‘trade off’ between the two.

Pollard (2014) utilises the broader term of ‘dependability’ which refers to the confidence placed in the assessment: *“it reflects the outcomes of the struggle to achieve validity and reliability”* (Pollard 2014: 385). Mansell et al. (2009) suggest that the notion of ‘dependability’ includes consideration of both ‘maximum validity’ and ‘optimal reliability’ (p12). The term ‘dependability’ is useful, but appears to mask the ‘trade off’ between validity and reliability, which is a key area for exploration in this study, thus the underlying concepts will remain part of this discussion.

Lum (2015) describes a recent paradigm shift in assessment, from a structure which aimed to compare pupils for the purposes of school or professional selection, to a structure where the focus is now on using assessment to support learning. A prime concern for 'assessment for selection' would be in making a highly reliable comparison, which would also be useful for school accountability measures; whilst assessment to support learning may be more concerned with a highly valid sampling of the construct. The different purposes appear to be requiring a different 'balance' in terms of validity and reliability, which may explain some of the tensions in the current system.

With large-scale collection of evidence and effective moderation procedures, where teachers compare and discuss judgements, Harlen (2007) argues that reliability of summative teacher assessment can be as high as it needs to be: 'reliable enough' to merit the conclusions drawn from them, 'reliable enough' for their purpose (Newton 2009). Many argue that teacher assessment is preferable to repeatable tests which narrow the curriculum (Wiliam 2003), and signifies a balance between the demands of reliability and construct validity. Moderation is hailed as: "*potentially the most effective strategy for ensuring both validity and reliability in teacher assessment*" (Johnson 2013: 99), supporting both consistency of judgement and teacher understanding of the breadth of the domain. Nevertheless, concerns regarding reliability of teacher assessment persist: "*the accountability function impedes the ability to use assessment as an integral part of the learning process, placing the teacher in a conflicted position*" (Green and Oates 2009: 233). It appears that moderation is a key area for this research to explore, in terms of increasing reliability of teacher assessment and the effect it has on teacher understanding of assessment processes.

2.3 Formative and summative assessment

2.3.1 Distinctions between formative and summative assessment

The distinctions between formative and summative purposes of assessment have received much attention in the UK during the last 20 years (Wiliam 2011), with authors asserting the importance of the former and utilising a number of names to emphasise the definitions. Following Black and Wiliam's review of assessment research (1998), the Assessment Reform

Group (ARG) argued for a clear distinction to be made between ‘Assessment **of** Learning’ (AoL), for the purposes of grading and reporting, and ‘Assessment **for** Learning’ (AfL), for the purpose of supporting learning (ARG 1999). The new terminology represented a call for: *“different priorities, new procedures and a new commitment”*, after: *“too much attention being given to finding reliable ways of comparing children, teachers and schools”* (ARG 1999: 2). By utilising a new term which contained ‘learning’, the aim was to promote a renewed focus on formative assessment which was felt to be the ‘key’ to improved learning (ARG 1999). Such a ‘rebranding’ was about shifting practitioner and policy focus, rather than the creation of new assessment concepts; the terms formative/AfL and summative/AoL are largely used interchangeably.

AoL aims to summarise pupils’ learning at a particular point in time for the purpose of accountability (Mawby and Dunne 2012: 139), for example, pupils are accountable for their performance towards certification or teachers are accountable for the performance of their pupils (Brown 2004). Such summaries of learning - either grades or narratives - can be reported, for example, to parents, other teachers, school leadership teams or school inspectors. In contrast, AfL is an: *“ongoing planned process that focuses on identifying the next steps for improvement”* (Harrison and Howard 2009: 28), the process is seen as an integral part of teaching and learning, providing feedback for both the teacher and the pupils. AfL requires the active involvement of children, and researchers stress the importance of dialogue and questioning (Black and Harrison 2004). Black and Wiliam suggest that: *“assessment provides information to be used as feedback... Such assessment becomes ‘formative assessment’ when the evidence is actually used to adapt the teaching work to meet the needs”* (1998: 2), thus it is the use of assessment information to support the learning process which distinguishes formative and summative assessment, rather than the assessment task itself.

Harrison and Howard (2009) suggest that AfL guidance, with its aim of promoting learning, can be applied more widely than AoL guidance, since summative assessment practices may be more defined by country-specific guidelines, whilst formative assessment focuses on more generic principles of teaching such as the importance of rich dialogue and identifying the learner’s starting point. Such widespread application perhaps explains the wealth of

research into formative assessment; however, changes to government guidance with regards to teacher summative assessment in primary science provides the opportunity for this study to provide new insights into the nature of summative assessment, in particular which information is used to make a summary judgement.

In recent years mounting evidence for positive impact of formative assessment on children's learning (e.g. Hattie 2009, 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 (William 2003) has led some teachers to view AfL and AoL as the 'good' and 'bad' sides of assessment respectively (Harlen 2013). However, there is also evidence that some teachers in the UK were misinterpreting AfL to mean frequent testing, demonstrating a lack of understanding of the aims of assessment practices (Black 2012). Repeated summative assessments could be perceived as a type of formative assessment if the learner receives feedback, although if the feedback is a numerical score or similar, then it would effectively end the exchange rather than open up the dialogue (Webb and Jones 2009: 173). Swaffield (2011: 433) also questions whether AfL and formative assessment are synonymous in practice, noting the: 'distorted practices that are erroneously termed AfL' in government policy (DCSF 2008). It appears that a simple split between formative and summative assessment has not led to universal understanding, making this a key area to explore in this research.

Some authors have argued that formative and summative are not separate forms of assessment, noting that it is: 'difficult to draw clear distinctions' (Davies et al. 2012), with the same tasks being used for both summative and formative purposes (Hodgson and Pyle 2010), for example, the formative use of summative tests (Black et al. 2003). Harlen (2007) suggests AfL and AoL differ only in purpose and degree of formality, whilst practitioners are likely to focus on the timing, with summative assessment coming at the end of a unit (Mawby and Dunne 2012). Authors suggest that rather than a dichotomy, it may be more useful to see these assessment processes as dimensions (Harlen 2013) or perhaps a continuum (William and Black 1996), which could be a useful line of enquiry for this research. Taras (2005) goes further and questions the underlying distinction arguing that: *"all assessment begins with summative assessment (which is a judgement) and that*

formative assessment is in fact summative assessment plus feedback which is used by the learner” (p466). Citing Scriven (1967), who was the first to use the terms formative and summative, Taras asserts that the process of assessment is the same, and the *“choice of function should not impinge on the actual process of assessment”* (2005: 468). In fact, she claims that the separation of formative and summative assessment is damaging since it requires separate systems of assessment leading to needless repetition (Taras 2007). The way teachers conceive and enact the relationship between formative and summative assessment is an area for research in this study.

2.3.2 A ‘formative to summative’ approach to teacher assessment

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 for change in assessment practices suggest that it is possible and desirable to use the same evidence for both formative and summative purposes in a system of ‘formative to summative’ assessment (Nuffield Foundation 2012). The ‘day-to-day, often informal, assessments’ (Mansell et al 2009: 9) which are used to inform next steps in learning, can be summarised at a later date. This does not mean, for example, comments for improvement should be accompanied by a summative score, since the comments are likely to be ignored if there is also a score (William 2011). However, if the evidence compiled from everyday interactions in the classrooms can be aggregated into a summary statement then the negative impact of summative testing could be avoided.

A similar suggestion was made in the Task Group on Assessment and Testing (TGAT) report (DES 1988), where it was suggested that: *‘it is possible to build up a comprehensive picture of the overall achievements of a pupil by aggregating, in a structured way, the separate results of a set of assessments designed to serve formative purposes’* (p4). Thus assessments designed for primarily formative purposes could also be utilised for summative purposes, although the reverse was deemed to be less practical, since summative assessments would take place at the end of a period of learning. It is interesting to note that when TGAT concluded: *“We judge therefore that an assessment system designed for formative purposes can meet all the needs of national assessment at ages before 16”* (DES

1988: 4), it was impossible to predict the changes in accountability measures to come. This study should consider how such a process of 'formative to summative' could function in the current context.

Nevertheless, there is not universal agreement that a 'formative to summative' approach, whereby assessment evidence is collected for both formative and summative purposes, 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), since the priority may be on judgements of independent performance, rather than to support learning. Stobart (2012) asserts that validity is principally tied to the purpose of the assessment, thus raising the question of whether validity is compromised if the same information is used for multiple purposes. Mansell et al. (2009: 7) also suggest that assessments designed for one purpose may not be fit for another, however, the concern over multiple use of assessment appears to be largely around its use for accountability: *"using assessment data for institutional monitoring can have a negative impact upon the quality of education in that institution, which clashes with the most fundamental of uses of assessment data in improving pupils' learning"* (p8). Thus it may be that utilising assessment data collected formatively to inform summative assessments is acceptable as long as the summative assessments are not 'high stakes'; perhaps the fact that science scores do not feature in English school accountability measures may provide an advantage here, but this also means that a system of 'formative to summative' teacher assessment may not be transferable to the 'high stakes' subjects of English and maths.

Wiliam and Black (1996) argue that all assessments have the potential to produce 'interpretable evidence' and it is possible to use assessment information for different purposes, as long as the elicitation of evidence is separated from the interpretation or judgement. Harlen (2007) 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). This appears to bring us back to a separation of formative and summative assessment, where the functions separate after elicitation. However, there also appears to be a suggestion that a 'formative to summative' approach may depend on the collection of evidence during formative assessment, perhaps devaluing more ephemeral learning experiences, like

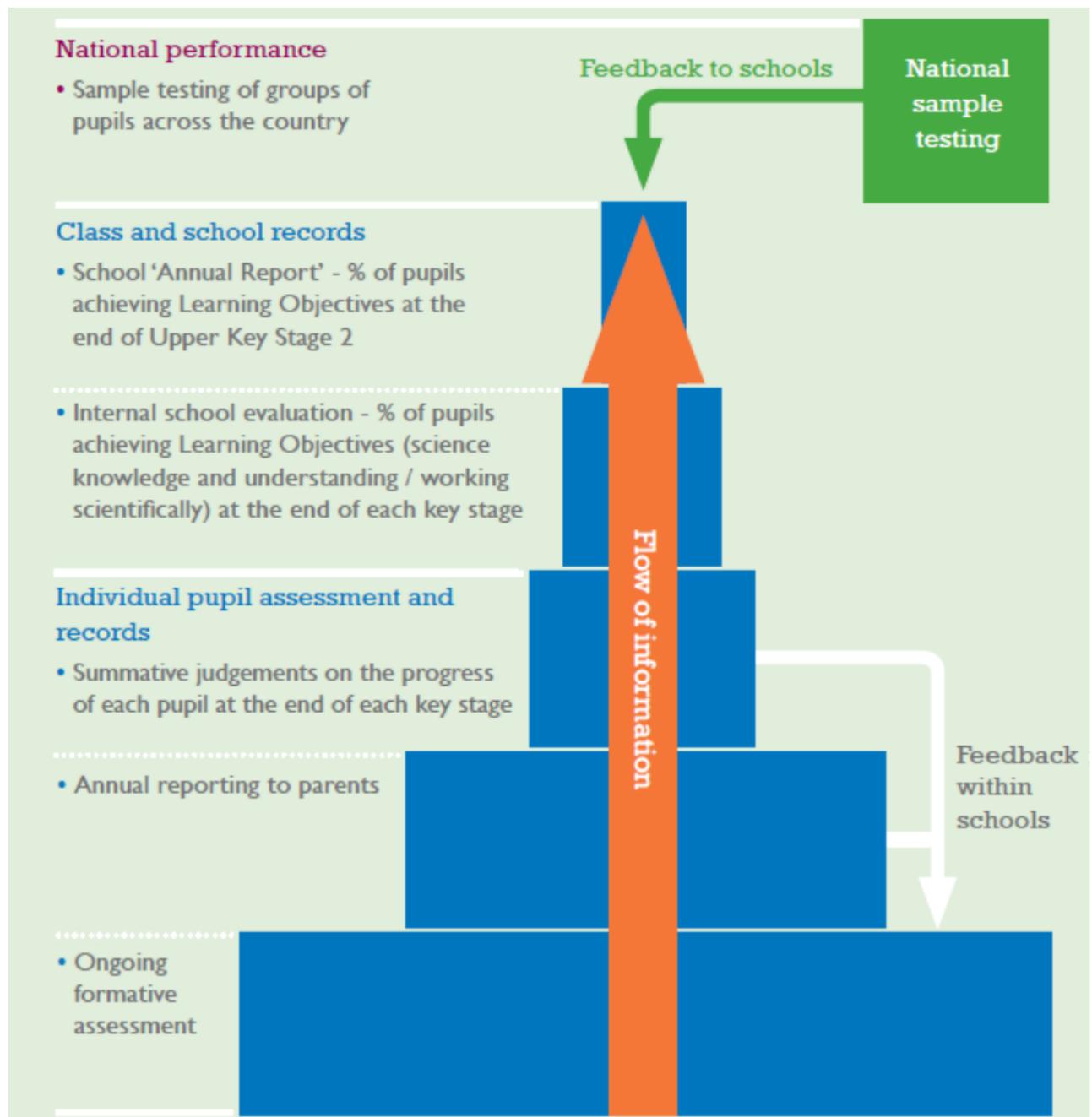
unrecorded discussions and explorations. Such a focus on recordable evidence is also reminiscent of the frequent summative assessment which has been described above as a misinterpretation of AfL (Black 2012). The collection of evidence could shift the focus onto the pupil's outcomes rather than the pupil's learning, assessment as an event rather than a process (Swaffield 2011). Klenowski 's (2009: 264) definition of AfL puts emphasis on the ongoing nature of assessment: *“Assessment for Learning is part of everyday practice by students, teachers and peers that seeks, reflects upon and responds to information from dialogue, demonstration and observation in ways that enhance ongoing learning”*. Whilst James et al. (2007) suggest that a guiding principle is to focus on learning rather than performance, akin to Davis' (1998) description of thin performance over rich knowledge. Harrison and Howard (2010) suggest that it is the balance between formative and summative assessment which is the problem. It is not clear from this discussion whether a focus on formative assessment evidence to inform summative assessment, as suggested by Harlen (2007), will lead to frequent summative assessment rather than a primacy of AfL, something which will need to be explored in practice during this study. 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, and this study will explore ways to do this.

2.3.3 A model of 'formative to summative' teacher assessment

The lack of centralised guidance for English primary teachers for how to assess science, particularly since the introduction of revised National Curriculum (DfE 2013a) and changes to assessment arrangements in 2013, means that there is: 'no single approach to teacher assessment' (Harlen 2012: 137) and researchers note the 'formidable challenge' (Black 2012: 131) of developing classroom assessment practices. In an attempt to fill the guidance vacuum, the Nuffield Foundation convened a group of experts, led by Wynne Harlen, who produced a pyramid-shaped 'formative to summative' model for teacher assessment (Figure 2.1, Nuffield 2012). This represents the drawing together of decades of assessment research, to both assert the importance of formative assessment, whilst providing a structure for drawing together summative assessment. This landmark 'formative to

summative' model provides a theoretical framework which can be tested and developed in practice.

Figure 2.1 The Nuffield data-flow pyramid model (Nuffield 2012: 20)



The base of the Nuffield pyramid (2012) represented the wide array of ongoing assessment practices which take place in the classroom. In order to summarise pupil attainment, some of this information 'flows' to the upper levels of the pyramid, for reporting to different groups. A key idea is that the assessment is based on classroom practice, but there is a reduction in the breadth and detail of data passed up at each layer, from 'rich formative assessment' to 'succinct summative information' (Nuffield 2012: 18). Also, rather than a national system of standardised testing for all, as happened with Key Stage 2 science SATs, the expert group recommended a 'national sampling' to satisfy monitoring requirements. This research will not consider the national sampling; instead it aims to provide a close examination of how a 'formative to summative' model could function in practice.

The Teacher Assessment in Primary Science (TAPS) project operationalised the Nuffield pyramid by working with teachers to consider assessment at each layer of the pyramid. The first published version of the TAPS pyramid school self-evaluation tool (Figure 2.2, Davies et al. 2014) re-ordered the upper pyramid layers so that summative judgements were made in a 'monitoring layer' before the reporting layer; formative assessment at the base of the pyramid was also split into teacher and pupil layers. The addition of a list of examples into each box was found to support practitioner interpretation of the statements (Davies et al. 2017) and this later developed into an interactive pdf with clickable boxes linked to examples from a range of classrooms and schools (Earle et al. 2015b). A later iteration of the TAPS pyramid (Figure 2.3, Earle et al. 2015a) included a 'shared understanding' box in the centre of the pyramid, to emphasise the importance of science and assessment literacy (Earle 2015, Davies et al. 2017).

Figure 2.2 TAPS pyramid school self-evaluation tool (Davies et al. 2014)

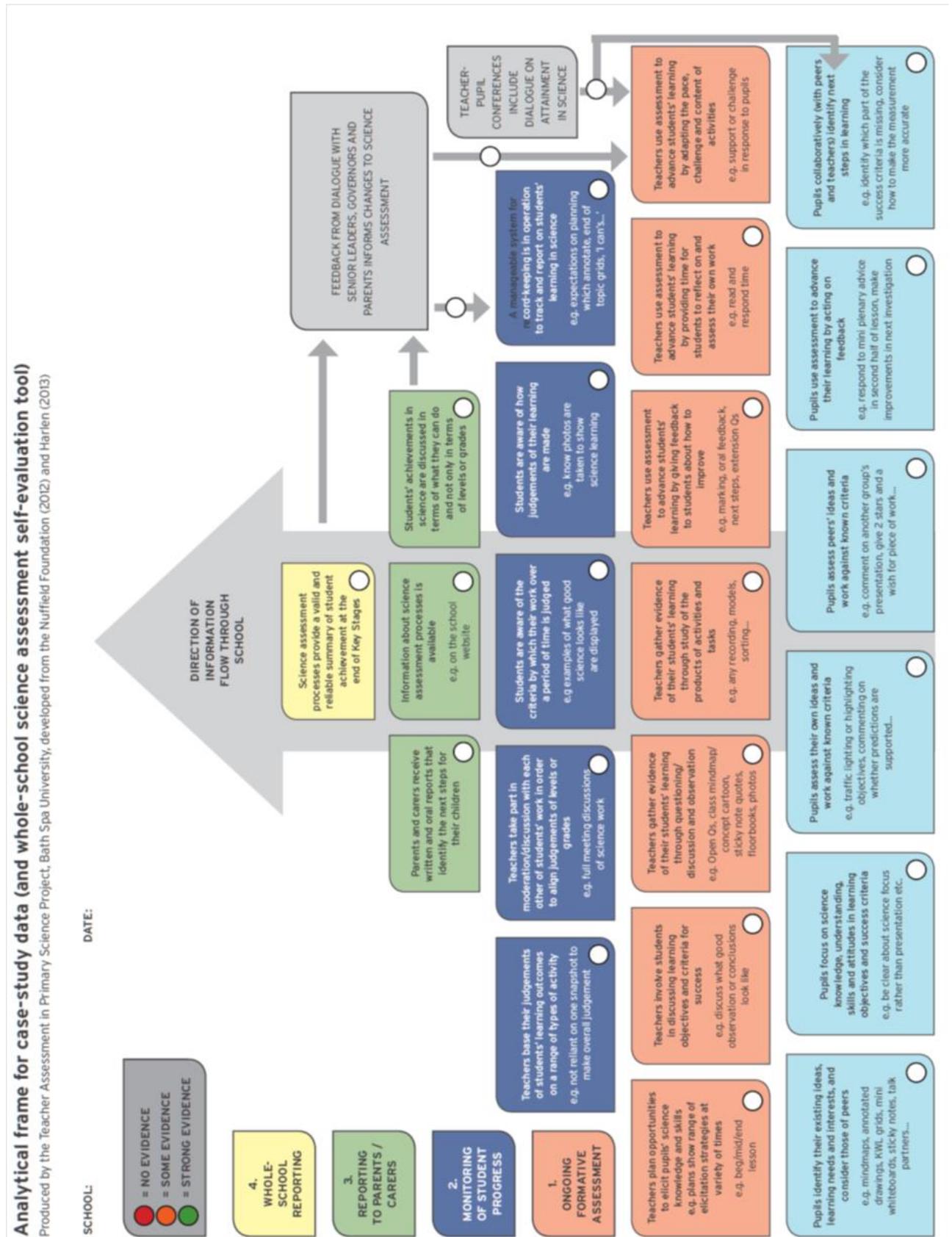
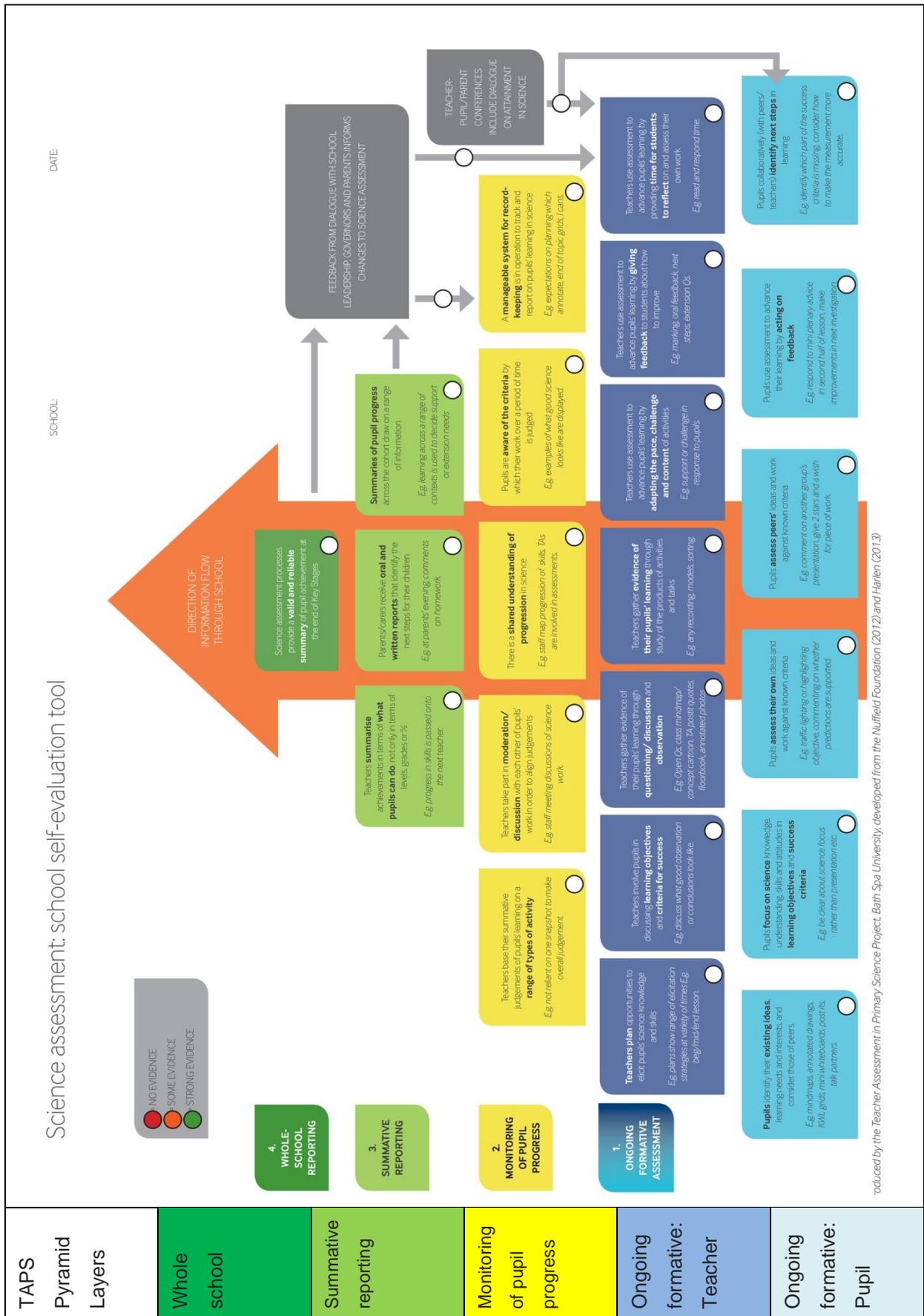


Figure 2.3 TAPS pyramid school self-evaluation tool (Earle et al. 2015a)



The TAPS pyramid includes the 'formative to summative' flow of information, which reduces in detail as it is summarised, which was validated as maintaining the Nuffield pyramid principles by a reconvened panel of the Nuffield group (Davies et al. 2017). However, whilst formative assessment is described in detail in the pupil and teacher layers, and summative reporting is detailed in the upper layers, the formative to summative transition is less clear, a view also expressed by members of the validation panel (Davies et al. 2017). The detail of the processes which support a move from 'formative to summative' use of assessment information provides the prime focus for this research and it is the area where new guidance for practice can be developed.

2.4 The nature of primary science education in England

2.4.1 Purposes and nature of science education

Before deciding whether an assessment is valid, it is necessary to understand the domain of the assessment, to consider what is being assessed, thus this section considers the nature of primary science.

The nature of primary science is in many ways determined by views regarding the purpose of teaching science. For some, the main aim of school science is to prepare future scientists, supplying the 'pipeline' (CBI 2015). However, this rather narrow view implies that science in the primary school will only be relevant for a minority of pupils, in the same way as saying the purpose of teaching of writing is to produce the next generation of authors. There is a value in understanding and interacting with the world beyond the production of future scientists, both to develop ideas and appreciation of the world and to be able to: *"engage effectively with different aspects of modern life"* (Harlen 2008: 12). A broader view of school science describes its aims in terms of 'scientific literacy' which proposes that engagement with science is necessary for all, since science permeates life choices: *"the ability to recognise and use evidence in making decisions as informed citizens"* (Harlen and Qualter 2014: 18). Primary science can provide the foundation for scientific literacy, and recent reports emphasise the need to start early with, for example, building 'science capital' at both home and school (ASPIRES 2013). Primary science for scientific literacy is a vision for

those in science education, but it cannot be assumed that this matches its perception in schools, since most primary teachers are not science specialists (to be discussed further in Section 2.4.3).

Black et al. (2002) found differences in the application of formative assessment in different subjects. Science is different from other subjects because children may come to the subject with strongly held pre-conceptions (Driver et al. 1985). For example, the Science Processes and Concepts Exploration (SPACE) project found that children develop their own ideas about the world (Russell et al. 1993), whether or not they are taught science, and that: *“without intervention to introduce a scientific approach in their exploration, many of the ideas they develop are non-scientific and may obstruct later learning”* (Harlen 2008: 9). Black et al. (2002) assert that the ‘mere presentation of the correct view’ has been found to be ineffective in addressing children’s alternative conceptions, and that discussion, challenge, evidence and argument are required (p16).

Science for young children is often linked to a constructivist, child-centred approach which champions play and practical experimentation (Pollard 2014). Piaget’s (1961) emphasis on physical exploration supports the value of inquiry-based pupil activities like science investigations. With the child leading the learning there appears a role for self-assessment, a key feature of Assessment for Learning (Black et al. 2003). Clarke (2001) asserts that children must be trained to self-evaluate, so that they can become more independent of the teacher and begin to monitor their own progress. This requires both an ability to step back, to judge their own performance, and knowledge of the standard or criteria.

In a similar way that the purposes of science education can be described in narrow or broad terms, the same could be said of the nature of science itself. A narrow view of science would be to describe it as a body of knowledge or ‘catalogue of facts’, or alternatively as a process of ‘child-led discovery’; these extremes do not represent the full nature of science (Dunne and Peacock 2015; Davies and McMahon 2011), such a process-content debate will be considered further in Section 2.4.3. Science: *“provides a way of making sense of the world”* (Howe et al. 2009: 2), building on children’s ideas towards ‘bigger’ ideas and the ‘big ideas’ of science (Harlen et al. 2010, 2015). In order to make sense of the world, the

individual needs appropriate attitudes, skills, knowledge and understanding: they need to be curious or interested in the world, be able to observe closely, and be able to link what they see to what they already know, thus science is both a body of knowledge and a process of discovery.

2.4.2 Inquiry terminology

Many labels are used to describe investigative science where pupils interact with materials and process the information gained, for example, 'process skills' (Harlen 1999), 'procedural understanding' (Roberts and Gott 2006), 'scientific enquiry' (DfEE 1999), 'Working Scientifically' (DfE 2013a) and 'inquiry skills' (Harlen and Qualter 2014). Whilst there are differences in terms of emphasis and listing of specific skills, they broadly all relate to: *"identifying investigable questions, designing investigations, obtaining evidence, interpreting evidence in terms of the question addressed in the inquiry, and communicating the investigation process"* (Harlen 1999: 129). Thus in broad terms, the choice of terminology used in this thesis, was driven towards the use of 'inquiry skills', in order that this research study has the widest potential audience, since this is the internationally accepted term seen in Inquiry-Based Science Education (see below). However, for the school-based sections the term 'Working Scientifically' will also be used since it refers to the particular curricular criteria used by the teachers; thus 'Working Scientifically' refers to a particular set of criteria, whilst 'inquiry skills' refers to the domain more broadly.

Dunne and Peacock (2015) suggest that it is vital that Inquiry Based Science Education (IBSE) should be incorporated: *"from earliest stages if the habit of enquiry is to be established"* (p23). IBSE is an approach to teaching which champions the inquiry process, from raising questions to planning, investigating and drawing conclusions (Harlen and Qualter 2014). This is not inquiry in isolation, but combines the development of both ideas and inquiry skills, as children: *"progressively develop key science ideas through learning how to investigate and build their own knowledge and understanding of the world"* (Harlen 2018: 37).

Nevertheless, much recent primary science assessment research has been concerned with the development of science concepts rather than inquiry skills (Hodgson and Pyle 2010, Black and Harrison 2004:18) and when skills have been addressed they are considered separately to concepts (e.g. Russell and Harlen 1990). However, the diversity of skills within the subject mean that the: *“assessment capabilities required by science teachers are wide ranging and complex”* (Edwards 2013: 212). Recent focus on the ‘Big Ideas’ of science contains recognition of the growth of IBSE and the importance of inquiry-based pedagogy, but the fourteen ‘Big Ideas’ focus on ideas of science and about science (Harlen et al. 2010, 2015). The importance of pupil talk and effective questioning to support formative assessment has been well documented (Mortimer and Scott 2003), but again it is the development of science concepts which dominate (Earle and Serret 2012). Therefore, a particular interest in this research will be to consider the assessment of inquiry skills, which are both ‘technically difficult’ and inhibited by ‘a content-dominated view of science education’ (Harlen 1999). It will be important to consider in practice how a teacher can assess scientific inquiry skills in action since the ‘teacher cannot be everywhere at once’ and ‘such learning is not always captured in the write up of the investigation’ (Davies et al. 2012: 248).

Despite a general consensus in broad terms regarding the nature of scientific inquiry, there is no definitive list of inquiry skills or inquiry types; they are ‘not well defined constructs’ (Millar 2010: 127), which poses potential difficulties when it comes to assessment, since there is a lack of agreement regarding the scope and criteria. An ‘ill-defined construct’ is difficult to operationalise in assessment terms; it is difficult to set assessment criteria for achievement of something that cannot be precisely defined. A shared understanding of inquiry across the science community is important for ‘adequate reliability’ in assessment (Halliday 2010). This study should explore how teachers enact assessment of inquiry skills.

2.4.3 Atomism and holism in the process-content debate

The ‘process-content’ debate concerns both the relative importance of skills and concepts within a subject, together with the relationship between the two, with implications regarding whether each should be considered separately in an atomistic way, or as a whole,

holistically. Wenham and Ovens (2010) describe three kinds of knowledge: “*know that, know why, know how*” (p10), which could be equated with science content, attitudes and skills. Knight et al. (2014) draw attention to the importance of the relationship between pedagogy, assessment and epistemology, regarding the nature of knowledge and what it means to know something as an integral part of the triad. Some separate ‘knowledge’ which is seen as factual information, and ‘understanding’ which is linked more with explanation, criticising that the drilling of facts does not lead to connected in-depth understanding (Davis 1998). This is not to say that facts are not important, but making links between the facts via thinking and experience is needed to develop learning for understanding (Harlen 2018: 33). Taking a pragmatic approach, the assessed curriculum for English schools lists concepts and skills, and thus both conceptual content (inclusive of knowledge and understanding) and inquiry skills will be discussed in this study, in order that guidance directly relevant to practice can be produced.

An area of debate, particularly pertinent to a piece of research on assessment, is whether it is possible, or indeed advisable, to separate science into component parts. The most recent English National Curriculum (DfE 2013a) asserts:

*“Working Scientifically’ is described separately in the programme of study, but must **always** be taught through and clearly related to the teaching of substantive science content in the programme of study” (p6).*

The teaching and assessment of inquiry skills takes place in a context, so any inquiry will draw upon science conceptual content, for example, when making predictions or drawing conclusions. Standish (2007) questioned whether it is possible to teach transferable skills in isolation, whilst Millar (2010) asserts that skills are ‘strongly content dependent’. However, Millar (2010) also cites a number of instances where researchers have found that tests of procedural understanding or observations of student performance were found not to correlate with tests of conceptual understanding: which for some would suggest that the tests may be sampling different constructs. Alternatively, the content within the different assessments may affect performance, reinforcing Millar’s assertion noted above regarding content dependency. In terms of assessment, one implication could be that utilising a range

of tasks across a number of content areas could provide a more accurate picture of student performance. However, Halliday (2010) suggests that reliable assessment of such understanding or 'rich knowledge' is not possible because this involves application to new contexts in open ended tasks whose marking criteria will be open to interpretation.

Harlen (1999) argues that inquiry skills are 'inseparable in practice' (p129) from the conceptual context in which they are applied. The Assessment of Performance Unit (APU) found evidence of the 'interdependence' between inquiry skills and conceptual context, with content making 'a considerable difference to achievement' when for example, 'making observations, planning an investigation or interpreting results', together with the development of understanding being dependent on the use of inquiry skills (Harlen 2008: 8). This means that any assessment of inquiry skills will be a combination of both the ability to use the skill and familiarity or knowledge of the content domain. Since science inquiry skills are so context dependent, a single assessment activity cannot reliably assess the use of individual skills; thus assessment of inquiry skills may only be possible by utilising information from multiple assessment opportunities.

Dunne and Maklad (2015) note a 'fuzzy' relationship between the practical doing of science and conceptual development, with a 'hands on' approach not necessarily being also a 'minds on' approach. Abrahams and Millar (2008) found that teachers appeared to separate the teaching of concepts and skills in their thinking and planning, and in practical lessons tended to focus on 'producing the phenomenon'. Eady (2008) found that in primary schools it was assumed that if pupils were engaged in practical activity then they would learn something. There appear to be a number of issues here which can be taken forward to explore in this research, for example, whether teachers are teaching and assessing concepts and skills separately, and if so, how they are doing this if it is not possible to teach inquiry skills without a conceptual context. The issue of whether it is advisable to separate science into its component parts is part of a wider debate of atomistic versus holistic teaching and assessment.

An atomistic approach is commonplace in schools, where the curriculum is separated into lesson-sized chunks, supporting teaching and assessment, particularly in terms of

manageability. The breaking down of inquiry into smaller skills was proposed by the AKSIS project and others as a means of direct and explicit teaching of particular skills (Goldsworthy et al. 2000, Coates and Wilson 2003). Such tightly defined learning objectives support atomistic assessment, where a different narrow focus is assessed each time. However, Sizmur and Sainsbury (1997) question whether atomised tick lists of behaviours present a view of the curriculum which is too narrow, whilst Swaffield (2011) notes that such convergent assessment has the potential of 'criteria compliance', where the focus is on the assessment rather than the learning. Lum (2015) describes prescriptive and expansive assessment, where 'prescriptive' assessment is based on predetermined outcomes, whilst 'expansive' assessment includes the idea that teachers make judgements about both the value of any one assessment and how any one piece of evidence sits in relation to others, thus 'expansive' assessment is more holistic.

Ollerenshaw and Ritchie (1993) argue for a holistic view of primary science, suggesting that practitioners should be: "*wary of fragmenting children's learning in science into arbitrary compartmentalised skills*" (p150). Since scientific inquiry is a continuous and complex process, it is difficult to segregate it into component skills. Critics of an atomistic approach highlight implications for both teacher and pupil: the teacher may be assessing skills out of context, and the pupil may lose the meaning of such skills and their place within a scientific inquiry. Digby (2014) suggests that holistic documentation of learning stories supports the learning process, avoiding misinterpretation through decontextualisation. This links to the previous discussion that skills are developed within a conceptual context, which needs to be taken into account during assessment.

Harlen (2006) suggests that any description of separate skills is a: "*convenience rather than an attempt to describe reality... We look at the components so as to help children develop skill in all aspects of enquiry*" (p96). McMahon and Davies (2003) suggest that a 'focused teaching' model could: '*bridge the gap between atomism and holism*' (p37), with specific teaching for component skills, which are then applied in the context of a real investigation. Once again a range of issues are arising for the research here, notably, whether an atomistic approach to the teaching and assessing of skills can be balanced with a holistic view of the nature of science.

2.4.4 Professional learning for teacher assessment literacy and subject leadership

An integral part of teacher assessment concerns teacher understanding of assessment: 'assessment literacy' (Edwards 2013), which will be a key line of enquiry within this study. Klenowski and Wyatt-Smith (2014: 2) assert that assessment literacy includes the ability to design quality assessments, as well as the ability to use criteria and evidence to make judgements. They go on to describe assessment as a 'shared enterprise', with teachers having a central role in assessment reform (Klenowski and Wyatt-Smith 2014). Connolly et al. (2012), also in Queensland where there is a long history of consensus moderation, describe how teachers share a 'common language' and are able to draw on multiple sources of evidence when making judgements. 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).

DeLuca and Johnson (2017) assert that despite widespread recognition of the need for assessment literate teachers, research has indicated a low level of assessment knowledge and skills in the teaching profession (p121). Black et al. (2011) found that teachers needed to first recognise that change in assessment practice was necessary; this was accomplished by considering the validity of current practices. They found that assessment competence involved a combination of literacy, skills and values (p452) and for this the development of moderation was key. Assessment practice will represent a complex mix of teacher understanding of both science and assessment, together with the teacher's perceived role in the school and of their own professional learning.

The use and development of formative and summative teacher assessment in primary science is dependent in part on the subject-specific professional learning of teachers within the school. In a primary school, the responsibility for subject development is likely to be shared across the school team, with each teacher leading for one or more subjects, perhaps those in which they have a special interest or expertise. There is no subject qualification requirement for subject leadership, therefore it is likely that many subject leaders (SL) will

be non-specialists, with the minimum science GCSE grade C (or equivalent) which is expected at the point of initial teacher training in England. Ofsted (2013) found in a survey of 91 primary schools that teacher subject knowledge was: *“not a serious barrier to pupils’ achievement”* (p12), with the fostering of enthusiasm and curiosity regarded as key features of successful primary science. Nevertheless, there will be a range of subject knowledge and confidence levels for science across the teaching staff, which has important implications for developments in teacher assessment, since it may be that teacher confidence in both the subject and the assessment processes need to be considered.

The term ‘leader’ is used in this thesis, rather than ‘coordinator’ since the latter implies a more managerial role, assisting with equipment for example, rather than strategic planning to move the subject forward (Bell and Ritchie 1999). A key role for subject leaders is to monitor what is happening across the school, since the science teaching would normally be carried out by the class teacher. This includes mapping the science content being taught across the school to ensure coverage and progression, which is particularly important in schools where there is topic-based teaching since it may not be clear where science is taking place (Harlen 2006). The management roles support the subject to happen, whilst monitoring tasks provide the SL with information to facilitate decisions for strategic direction and staff needs for professional development. Bianchi (2017) proposed a Trajectory of Professional Development, an arrow which described the way teachers moved through stages of pre-engagement, participation, collaboration and co-creation on their journey of professional learning. This has relevance for the teachers working on the TAPS project, who are the focus for the case studies in this research, since teachers may not immediately be at the ‘collaboration’ or ‘co-creation’ stage, perhaps needing time to ‘participate’ first.

The SL will need to balance the ‘multiple realities’ of the staff when implementing change in assessment practices, developing a clear vision which takes into account the ideas and experiences of all the people involved (Fullan 2016). During the PSQM process a whole school vision is developed in the form of a ‘Principles’ document, which *“provided a common understanding of what science in the school should look like and brought everyone together with a common agreed purpose and vision”* (White et al. 2016: 52). Porritt (2014)

describes the importance of collaboration, engagement, ownership and reflection for professional learning opportunities. Guskey (2002) argues that professional development first leads to changes in practice, and if these are successful, then teachers' attitudes and beliefs may change, but this process takes time and can be difficult for teachers. Porritt (2014) suggests that: 'putting knowledge to work' is an effective way of thinking about the impact of professional learning and development. This suggests that there may be a 'mismatch' between assessment rhetoric and assessment practice (Murphy 1999), conceptualisation and enactment of assessment may not necessarily be at the same stage, with either changes in practice leading to change in beliefs about assessment (Guskey 2002) or changes in assessment knowledge being 'put to work' (Porritt 2014).

2.5 Summary and research questions

This chapter has provided an overview of literature in the fields of teacher assessment and primary science education. Table 2.1 summarises key issues arising from the literature review, together with areas of focus for this study:

Table 2.1 Summary of issues and areas for focus

Section	Issues arising from literature review	Areas of focus for this study
2.2	<p>Purpose is key to assessment validity.</p> <p>Construct validity is a particular issue for primary science because of the difficulty of assessing practical activities.</p> <p>Teacher assessment is put forward as a more valid means of assessment, but concerns remain regarding its reliability, although moderation could support this.</p> <p>A possible 'trade off' between validity and reliability in teacher assessment.</p>	<p>Purposes of teacher assessment in primary science.</p> <p>Validity and reliability in teacher assessment, particularly for inquiry skills.</p> <p>The use of moderation could support teacher assessment in primary science.</p> <p>What a 'trade off' between validity and reliability could look like in action, in the current primary science context.</p>
2.3	<p>Distinctions are made between formative and summative assessment.</p> <p>Which types of assessment can fulfil which purposes.</p> <p>A 'formative to summative' model was proposed, but the process of moving from formative to summative purposes was unclear.</p>	<p>How distinctions between formative and summative are conceived and enacted in the current primary science context.</p> <p>What kind of assessment information could be used for formative and summative purposes.</p> <p>Analysis of processes involved in utilising formative assessment for summative purposes.</p> <p>Identification of the place and process of moving from formative to summative in a data-flow pyramid model.</p>
2.4	<p>There are debates in the relationship between inquiry skills and conceptual understanding, together with how atomistic or holistic assessment should be.</p> <p>Changes in assessment practice may be led by science subject leaders, but teacher professional learning is tied up with both subject understanding and assessment literacy.</p>	<p>Atomistic and holistic teacher assessment of scientific inquiry and concepts.</p> <p>Changes in assessment literacy and practice over time.</p>

This study seeks to develop understanding of the relationship between formative and summative assessment in primary science, in order to inform guidance for practice in teacher assessment in primary science. The literature review has identified that the purposes and processes for making teacher assessments, and relationship between formative and summative assessment are unclear in the current context, thus the following research questions (RQs) are proposed:

RQ1. How do teachers assess children's learning in science for **formative and summative purposes**?

RQ2. How can teachers' conceptualisation and enactment of the **relationship between formative and summative** assessment of children's learning in science be used to inform guidance for practice?

RQ3. How can study of **changes over time** in conceptualisation and enactment of the relationship between formative and summative assessment be used to inform guidance for practice?

RQ1 is a scene-setting question, to explore current practice in English primary schools, which has not been mapped since the removal of standardised testing at age 11 in 2009. It is necessary to understand the current context in order to be able to provide relevant guidance to support practice. RQ2 considers the way teachers conceptualise the relationship between formative and summative assessment, together with the way this is enacted in the classroom. This question is designed to shed light on the processes of assessment, in order to identify principles to support practice. Whilst RQ3 examines how such conceptions and practice change over time in response to developments, considering forms of guidance which may have an impact and the factors affecting the processes of change in assessment practice over time. The next chapter will explore the methods which will be used to address the three RQs.

Chapter 3 Methodology

3.1 Introduction

The aim of this research is to develop understanding of the relationship between formative and summative assessment in action, in order to inform guidance for practice in teacher assessment in primary science. This aim places the research within an 'Integrated Knowledge Tradition', bringing the practical knowledge of 'knowing how' together with the theoretical knowledge of 'knowing that' (Furlong and Whitty 2017). This chapter will explore the methodological considerations of such empirical enquiry which engages with real world settings to develop both theory and practice.

In response to the gaps identified in the literature review regarding understanding of the relationship between formative and summative assessment in action, the following research questions (RQs) were proposed:

RQ1. How do teachers assess children's learning in science for **formative and summative purposes**?

RQ2. How can teachers' conceptualisation and enactment of the **relationship between formative and summative** assessment of children's learning in science be used to inform guidance for practice?

RQ3. How can study of **changes over time** in conceptualisation and enactment of the relationship between formative and summative assessment be used to inform guidance for practice?

In order to answer the three research questions, an applied research approach was needed to produce guidance and evidence-informed principles which would be directly relevant to practice. By working in collaboration with practitioners to trial and improve practical guidance, the aim was to develop both theoretical and practical products. A Design-Based Research (DBR) approach was selected because it matched the aims of the research, since its key features, which will be explored further in Section 3.2, include: collaboration with

practitioners to meet dual goals of developing both theory and practice through iterative cycles in real contexts.

This chapter will begin with consideration of a Design-Based Research approach, before outlining the sample, methods, ethics and analysis, together with consideration of validity and reliability in such social science research.

3.2 Design-Based Research

3.2.1 The nature of Design-Based Research in this study

Educational research has been increasingly criticised for a lack of impact on practice, which could be interpreted as a failure of consequential validity (Hartas 2010). The 'Integrated Knowledge Tradition' has developed in response to this, with the aim of bringing theory and practice closer together (Furlong and Whitty 2017). Design-Based Research is a methodology within this tradition; it is interventionist research which aims to engineer products and develop recommendations to inform practice and support educational reform (Brown 1992: 143). Before moving on to the practical features of DBR, this section will consider the place of DBR within educational research.

Any research is underpinned by beliefs about the nature of reality and what it is possible to know, placing it within a research paradigm, the accepted way of thinking which underlies the research. DBR can be aligned with both positivist and interpretivist paradigms; each of which will be explored in turn in order to clarify the nature of this research study. Firstly, positivist ontology, regarding the nature of reality, asserts the existence of objective truths (Hitchcock and Hughes 1995: 22). Positivist epistemology, regarding the nature of knowledge, proposes that researchers can discover causal links by measuring effects of actions in controlled situations, separating the observer from the observed (Lincoln and Guba 1985). By simplifying social reality into cause and effect relationships, positivists can put forward generalisations, for example, to improve standards in schools. Design-Based Researchers who utilise Randomised Control Trials could align to a positivist paradigm, with a 'Learning Sciences' view of educational research (Furlong and Whitty 2017). However, I would argue that it is the assumption that causal links can be examined separately from

context, which limits the application of positivist research in education: for every school, classroom and child has a unique set of circumstances that cannot be controlled or replicated. Simplifying the social situation can lead to inappropriate inferences that fail to consider the 'bigger picture' (Noyes 2004) and generalisations which are not helpful to the professional (Cohen et al. 2011).

At the opposite end of the paradigm spectrum, interpretative ontology assumes that each individual creates their own 'truth', social reality is subjective not objective. This research paradigm is not about the traditional cause and effect of the physical sciences, it is about rich descriptions of context and interpretations of the social world (Hitchcock and Hughes 1995). This results in the consideration of different viewpoints and layers of meaning within complex situations, studied within their natural setting: *"because realities are wholes that cannot be understood in isolation from their contexts"* (Lincoln and Guba 1985: 39). From an interpretivist perspective, socially constructed truths do not stand still for researchers to examine them; they are part of the situation at the time, and can be analysed from different viewpoints. The dynamic nature of social reality is a particularly salient feature of the Design-Based Research approach.

Interpretivist, or constructivist, researchers assert that there is not a universal truth to find in social science; reality is perceived by people in a particular context, it is experienced and socially constructed (Sprague 2010). My use of DBR is placed within the interpretivist paradigm since my ontological assumptions include a rejection of 'one truth' to 'solve' assessment in primary science, but there could be principles and exemplars from which others could extrapolate ideas to try in their own classrooms; rich description providing opportunities for transferability (Greene 2010).

This research sits within the interpretivist research paradigm since it places value on the perspective of the individual, providing insights into participant perspectives, layers of meaning and 'multiple ways of seeing' (Cresswell and Plano Clark 2011). Positivists assume that truth can be found by removing subjective judgements and interpretations, that good science can and should be value-free (Sprague 2010: 79). However, I would argue that data cannot speak for itself, the interpretivist researcher is the data gathering instrument and

forms part of the system being studied (Luttrell 2010). In this DBR approach both researcher and practitioner are active agents, co-researchers where each interaction is another source of data, and analysis aims to develop understanding of the processes involved in assessment in primary science in particular contexts.

3.2.2 Features of Design-Based Research

Design-Based Research is an emerging methodology in the field of education, with the following key features, each of which will be explored more fully below:

- Dual goal of product design and development of theory
- Iterative cycles in real contexts
- Collaboration between researchers and practitioners
- Use of a range of methods

In DBR the development of theory and products to support practice are intertwined (Design-Based Research Collective 2003). The aim or ‘design goal’ for this research was to do more than study the issue, it was to develop a better understanding of the relationship between formative and summative assessment in order to have an impact on practice. The Design-Based Research Collective (2003) asserts that research must lead to sharable theories that help communicate relevant implications to practitioners. The goal of refining a theoretical model for ‘formative to summative’ teacher assessment, defining design principles (Anderson and Shattuck 2012), was combined with the practical requirements of designing and testing ways for this to be implemented in real contexts.

DBR involves collaborative partnership between researchers and practitioners (Anderson & Shattuck 2012) in order to: “*generate evidence-based and ecologically valid recommendations for practice*” (McGuigan and Russell 2015: 35). The approach necessitates response to user feedback through iterative cycles of designing and testing; requiring the theory to do ‘real work’ in real contexts (Cobb et al. 2003). Such multiple iterative cycles can enable documentation of how designs function in authentic settings; including success, failures and interactions which refine our understanding (Design-Based Research Collective 2003).

Collins et al. (2004) note the importance of multiple ways of looking, in order to consider the many layers of the school learning environment, consequently, Design-Based Researchers typically use a range of methods (Anderson and Shattuck 2012). For some this leads to a 'Mixed Methods' approach to research, however, as an interpretivist I am interested in the qualitative understanding of participant perspectives and experiences of the relationship between formative and summative rather than the quantitative measurement of such viewpoints. Using quantification in qualitative data analysis is not the same thing as adopting a quantitative methodology, thus numerical summaries are consistent with this study. The choice of methods and means of analysis will be directed by the RQs and will be discussed further below.

Multiple levels of data can: *"result in greater understanding of the learning ecology ... of a complex interacting system"* (Cobb 2003: 9), but leads to challenging selection and analysis, particularly in DBR where the practice is not 'frozen' (Cohen et al. 2011). Trying to capture changing practices is also linked to the challenge of capturing practice in which the researcher plays an active part, a common criticism levelled at qualitative research, which is accentuated here by the collaboration seen in Design-Based Research (Anderson and Shattuck 2012). As discussed above in section 3.1, an interpretivist researcher is both part of the research context and acts as the data-gathering instrument (Lincoln and Guba 1985), plus within a DBR approach, an impact on practice is part of the aim. Researcher effects are not ignored, they are studied and analysed as an important part of the process. Shavelson et al. (2003) argue for a questioning approach to the role of the researcher in DBR, with the researcher taking a critical stance to their data and analysis, actively looking for alternative explanations and theories during the iterative and collaborative process. We will return to this discussion during consideration of validity and reliability in section 3.5.

3.2.3 Phases of Design-Based Research

There are a wide variety of models of DBR, for example, Herrington and Reeves (2011) define four phases: analysis/exploration, development of solutions, implementation/evaluation, and reflection (p597-598). Whilst Easterday et al. (2014)

propose six phases (Focus, Understand, Define, Conceive, Build, Test), each of which can have nested cycles within, although these are described in the largely positivist terms of testing an intervention using an RCT. Shah et al. (2015) provide a useful distinction between ‘macro- cycles’ which are focused on theoretical knowledge generation, and ‘micro-cycles’ which are more focused on generating ‘local practical knowledge’ regarding implementation in the context (p159). The DBR Phases utilised within this research pertain to ‘macro- cycles’, in line with Herrington and Reeves (2011), although their final phase of reflection is not listed as a DBR phase for this study because it took place during the writing up of the research, after the active collaboration with participants and iterative cycles had ended. However, this does indicate that further DBR research will be needed to continue to refine the products of this study, in partnership with practitioners.

The DBR phases in this research were named to link the phase to its stage in the theory building process: during the Exploration Phase the focus was on framing the issue; during the Development Phase the key focus was on developing and exemplifying ‘formative to summative’ assessment; in the Implementation Phase the use of ‘formative to summative’ assessment was explored over time. The DBR phases provided a ‘macro’ structure to the research process and supported comparison across time, pertinent to RQ3. A broad map of the DBR cycles or phases can be found in Table 3.1.

Table 3.1 Design-Based Research Phases mapped onto research questions and key data

DBR Phase	Outline of time	Research question focus	Key data
1. Exploration	March 2013 – February 2014	RQ1 Formative and summative assessment	Primary Science Quality Mark database (<i>March 2013</i>)
2. Development	February 2014 – March 2015	RQ2 Relationship between formative and summative	School A case study (<i>June 2013 – June 2015</i>)
3. Implementation	March 2015 – June 2016	RQ3 Relationship between formative and summative over time	School B case study (<i>March 2013–June 2016</i>)

Table 3.1 provides an outline mapping only because there is not a distinct halt to one DBR Phase before the next begins, however, the key events in the TAPS project impacted the

PhD study and provided useful markers for shifts in focus. For example, the first iteration of the TAPS pyramid self-evaluation tool was shared with project schools in February 2014 (Davies et al. 2014) and this marks the beginning of the 'Development' Phase. Table 3.1 includes key data to support an understanding of the DBR Phases, the selection of such data provides the focus for the next section.

3.3 Research sample

3.3.1 Sampling

My sample is drawn from the population of teachers in primary schools in England and the sampling techniques were driven by the RQs. This study is based on two kinds of sample: one set of submissions from a pre-existing database and two schools where practice was examined over a period of two or three years. This section will outline each sample, before moving on to a more in-depth discussion of the case studies, which form the larger proportion of the thesis.

In order to answer RQ1, regarding practice in science assessment, I utilised a pre-existing dataset from the Primary Science Quality Mark (PSQM). PSQM requires the science subject leader (SL) 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 13 PSQM 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 initial analysis in DBR Phase 1 since this could provide a 'snapshot' of approaches taken by English primary schools to the formative and summative assessment of pupils' learning in science.

The PSQM dataset allowed a sample of schools from across England to be utilised to answer RQ1, however, it could not be considered a representative sample since the schools were self-selecting, which could mean that their practices would be 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 likely that non-sampled schools may have had 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 was closely matched to RQ1, providing a framing of the issues for this study and for the beginning of the TAPS project, the first Phase of DBR. It also led to a realisation that RQ2 would need a different approach: "*a need exists because one data source may be insufficient*" (Cresswell and Plano Clark 2011: 8). The relationship between formative and summative assessment was not explicitly described in the C2 reflections; exploration of this relationship required interaction with participants.

In order to study the conceptualisation and enactment of 'formative to summative' assessment (RQ2) and its development over time (RQ3), there needed to be in-depth and ongoing analysis of practice in action. The use of case study within DBR will be discussed below, the focus here will be on the sampling of the schools. Two schools were selected as type exemplars, selected to illustrate what is possible (Newby 2010), a purposive or critical case sampling of schools (Teddlie and Yu 2007). The cases were not selected to be representative, but to be informative (Cohen et al. 2011), this purposive sampling was driven by the research questions which required exploration of change over time; the goal was depth rather than breadth of information-rich cases (Mears 2012). These are special cases, TAPS project schools for which I was the link tutor, enabling me to gather data about their assessment practices over time. Participation in an in-depth study over two or three years already suggests that these schools are atypical; such participation requires the support of the head teacher and the SL for repeated school visits and project days. Such commitment to remaining an active member of the project is likely to depend on science being given high priority in the school, so in effect I have only looked at schools where science is 'strong', but to answer my RQs I need the 'right source' (Newby 2010) which could commit to long term involvement.

3.3.2 Case study within DBR

In order to understand what is happening during DBR research when teachers try to use formative assessment information to make summative judgements, there needs to be a

detailed study of the process. Sufficient data is required to explore potential significant features of the case, there needs to be 'thick description' (Geertz 1973). Case study is in-depth study of a situation in its natural setting which is 'strong in reality' (Adelman et al. 1976: 148). For me this does not mean a single positivist objective reality waiting to be discovered, it sits within an interpretivist constructed reality where different people have different perceptions (Bassegy 1999), whether that be the teachers, the researcher or the readers of research. A case study includes: "*the study of an instance in action*" (Adelman et al. 1976: 141) and it is understanding of the 'in action' element which is so central to DBR and to this study of assessment.

A case study is the: "*study of singularity conducted in depth in natural settings*" (Bassegy 1999: 47), with a 'singularity' being the particular event, practice or situation under scrutiny. Adelman et al. (1976) call the case a 'bounded system' which is selected as an 'instance of a class', for example, a particular school as an instance of primary schools in England. There are two ways of setting up a case study, either a hypothesis leads to a case being selected, or a case leads to a hypothesis, the case study is a 'step towards theory' (Stake 2006). This research is an example of the former: the hypothesis that formative assessment could support summative assessment led to the identification of schools which were working towards this. Of course, behind the hypothesis were previous experiences with schools which helped develop the hypothesis, but this pre-dates this study. In this research the cases were chosen to support development of the understanding of the 'class' of assessment in primary science.

The iterative cycles of DBR can be analysed using case studies, with the role of the researcher acknowledged as a potential variable within the enquiry: by asking questions and observing they may change the situation they are studying (Bassegy 1999: 43). It could be argued that DBR goes further than noting the influence of the researcher, it is an explicit aim of DBR that the situation be developed in collaboration with the teachers. Case study can provide an in-depth consideration of the process: case studies are: "*a step to action. They begin in a world of action and contribute to it*" (Adelman et al. 1976: 148). Walker (1983: 163) notes that case study research may be accused of trying to 'embalm practices' which are always changing. Perhaps the iterative cycles of DBR can begin to address this by placing

a pause for reflection during each cycle. The changing practices of teachers within the study provide another layer for analysis, reinforcing the need for the depth of data collection which is an essential feature of a case study.

3.3.3 Case studies in this research

Stake (2006) suggests multiple case study research seeks to understand the ‘quintain’ (equivalent to ‘class’ above) by observing in multiple situations. The cases are studied for what they reveal about the quintain, however, Stake is keen to stress that multi-case research is still primarily concerned with the case; it is not a simple comparative study which looks for similarities and differences on a small number of attributes (Stake 2006: 82). I feel that my research does not meet Stake’s multi-case criteria, for although I am keen to understand the detail of what is happening when teachers assess, I am using case study as a method for testing and developing theory, rather than aiming to develop a better understanding of assessment in that particular school. Thus although multiple case studies are utilised in this research, each case is used for a different purpose, as shall be explored in the next section.

The case studies contained in this research focus on different periods within the DBR process and are presented for different purposes, perhaps making them different types of case study. Stake (2006) splits case studies into ‘intrinsic’, where the researcher is interested in the case for its own sake, and ‘instrumental’, where interest in the case is driven by an outside concern. Both of my case studies can be placed into Stake’s ‘**instrumental**’ case study category since they have been selected as ‘test-beds’ for the model of ‘formative to summative’ assessment proposed as part of the TAPS project. Yin (2014: 238) divides case studies into: ‘exploratory’, whose purpose is to identify research questions; ‘descriptive’, which describe the phenomenon in its real world context; and ‘**explanatory**’ whose purpose is to explain how or why a condition came to be. The last category appears to be the closest to my case studies as I am trying to explore how formative assessment information can be used for summative purposes.

Bassey (1999: 62) identifies ‘**theory seeking**’ and ‘**theory testing**’ case studies and it is on this categorisation which I can separate the two case studies in this research. Case A is ‘theory seeking’ since it is focused on the Development DBR phase, at the stage when theory is being developed. Case B is ‘theory testing’, since it is focused on the Implementation DBR phase where the TAPS pyramid model is being used in school. The iterative cycles of DBR call into question the separation of the cases in this way, since with each new piece of information theory seeking and testing could be taking place, however, by categorising the cases in this way I am providing a shorthand for their purpose rather than a full description of all elements.

DBR methodology requires that such case studies result in useful products for researchers and practitioners, which appears to be at odds with the typical aim of a case study where in-depth understanding of the case is the aim. If it is not possible to have freedom from time and context then it could be questioned whether it is possible to make generalisations; added to this: *“the trouble with generalisations is that they don’t apply to particulars”* (Lincoln and Guba 185: 110). Simons (1996) describes this as a paradox, focusing on the unique whilst seeking to generalise. She suggests that the research should challenge the reader to construct the generalisation for themselves: *“construct their own meanings from the evidence we offer”* (Simons 1996: 232). Rather than seeking statistically generalisable positivist ‘truths’, the reader applies the assertions to their own situations, making ‘naturalistic generalisations’: *“the responsibility for making generalisations should be more the reader’s than the writer’s”* (Stake 2006: 90).

The concern for passing the ‘responsibility’ to the reader is that the teachers and policy makers who are able to act on the lessons learned from a case study will not have the time or means to access the full reports. Bassey (1999) suggests ‘fuzzy generalisations’ can act as sound bites from research: *“Do y instead of x and your pupils may learn more”* (p51). Such summary statements provide accessible conclusions which are tentative, to acknowledge the many variables involved in classroom learning. Design-Based Research aims to support action from research and so it is part of my research design that such ‘fuzzy’ or ‘tentative generalisations’ will be included in the outcomes of the case studies, which can then be tested against other cases and by readers in other contexts.

3.4 Methods of data collection

3.4.1 Overview of data collection methods

This section will introduce the range of data collected and methods used to address the research questions. Table 3.2 provides an overview of data collection methods, with each method explored more fully below. For more detail on each item and links to secure electronic folders of the raw data, see Appendices 4A, 5A and 6A. School B data was collected over a longer period of time because the focus for this case was longitudinal change over time (RQ3), whilst for School A the focus was on the school's practice of 'formative to summative' (RQ2). Using a range of data from each school provided a range of lenses through which to explore the case, together with supporting the triangulation of data, as discussed further in Section 3.5.

Table 3.2 Overview of data collection methods

	PSQM database	School A June 2013 – June 2015	School B March 2013 – June 2016	Key data for RQ
Documentation <i>E.g. school documents collected on visits: policies, lesson plans, records, work samples, PSQM submission.</i>	91 assessment reflections <i>Items=91</i>	6 school visits 1 PSQM submission <i>Items=44</i>	6 school visits 2 PSQM submissions <i>Items=58</i>	RQ1 RQ3(SchB)
Non-participant observation <i>E.g. Lesson observation or observation of meeting/presentation</i>	-	4 lessons 1 meeting 1 presentation <i>Items=9</i>	3 lessons 1 meeting 2 presentations <i>Items=11</i>	RQ2 RQ3(SchB)
Semi-structured (researcher-led) discussions or meetings <i>E.g. interview/meeting, group discussion</i>	-	2 interviews 4 gp discussions <i>Items=6</i>	3 interviews 4 gp discussions <i>Items=4</i>	RQ2 RQ3(SchB)
Written tasks (researcher-led) <i>E.g. completion of questionnaire, sorting activity, pyramid self-evaluation on TAPS development days.</i>	-	6 development days <i>Items=8</i>	8 development days <i>Items=13</i>	RQ2 RQ3(SchB)
Total items in case record	91	67	86	

3.4.2 Documentary extracts from PSQM database

Extracts from the Primary Science Quality Mark (PSQM) database were selected to address RQ1 since they could provide information regarding assessment practices from across England. Round 4 submissions were the most recent submissions at the time, with schools working towards the Quality Mark in the year April 2012 to March 2013, and writing the final reflections in March 2013. Each teacher reflection consisted of around 200-400 words and described: practice within the school, changes across the PSQM year, their impact and possible next steps. The C2 reflection related to assessment practice so this was downloaded from the PSQM server and anonymised for each of the 91 English schools in Round 4. The rest of the submission would have provided a range of data regarding practices across the school, but it was only C2 which was purely focused on assessment, and so it was this reflection which was analysed.

All participating PSQM schools were informed that submissions may be used anonymously for research purposes and the most recent PSQM applications (Round 4) from all English schools (N=91) received an additional email regarding this study, providing them with the option to withdraw their data. Since the PSQM database was a pre-existing dataset, once access had been gained to the website, then data collection merely involved downloading the C2 reflections. Collection of data for the school case studies was much more varied and this will be the focus for the rest of this section.

3.4.3 Documentation and written tasks within case studies

Documentation is the written record of an event or process, on paper or electronically. These documents are relatively easy to collect, but may be selective in their representation of the context (Cohen et al. 2011: 236), together with only providing insight into the output rather than the process of construction. Nevertheless, the wide range of documentation noted in Table 3.2 is utilised in schools and can provide part of a rich bank of evidence. Collection of documentation largely took place on school visits. At the beginning of the TAPS project schools were given a list of assessment samples to provide, for example, policies, pupil assessments and tracking grids (for the full list see Appendix 3A). Photocopies or

photos of teacher planning and pupil work were also collected at each lesson observation. The documents were used to support discussions with the class teachers, and then scanned for inclusion in the electronic folder for each school. The aim was to collect a wide range of documents over time; although it should be noted that these would always be documents which were supplied by the SL or class teacher. This self-selection could provide a different picture to typical practice, for example, planning for lesson observations could be more detailed and pupil work may have been marked in a different way. However, this provides information about what the teachers think of as 'best practice', providing for another line of enquiry regarding teacher perception.

Both Schools A and B completed their own PSQM submissions during the case study period and this provided a wealth of data: action planning, SL reflections, CPD logs and a portfolio of evidence (in the form of a powerpoint). These documents were produced, largely by the SL, for the purpose of gaining an award, so they are again likely to represent what the SL viewed as 'best practice'.

Documentation was also produced as part of the TAPS project, including an application to join the project, together with a range of group and individual written tasks from the project development days which supplemented the school-led data above. For example, on TAPS development day 1 the teachers were asked to record individually what they understood by formative and summative assessment (Appendix 3B) and then to complete a group card sort of assessment strategies (Appendix 6D). These researcher-led tasks directed the participants to explore the language of assessment in a way that the school-led documents did not.

In addition, focused questioning in the form of short questionnaires was also used later in the project, to support the teachers to reflect on whether assessment practices had changed (Appendix 3C). Such questionnaires provided a range of largely open-ended questions in order that teachers had the freedom to explain their thoughts rather than assign them to a pre-determined category (Oppenheim 1992), since the aim was for in-depth study rather than comparisons across the population.

3.4.4 Observations within case studies

Observation within this research involved watching teachers: lead science lessons, take part in school staff meetings or present examples of their practice to a teacher audience. It was important to see assessment practice in action, whilst recognising that the act of observation will have an impact on the situation being observed (Bassey 1999). The practice will also be seen through one person's perspective; there is no such thing as an objective observer. All observation is subjective; the observer reconstructs their observations to create their own interpretation of the context (Greene 2010). However, post-lesson discussions with teachers were held to support the development and sharing of interpretations of the observations.

In order to examine practice in action, a number of lessons were observed. This provided the opportunity to explore in-class processes, together with the relationship between reported and actual assessment behaviours, the ideal and the real (Angrosino 2012). The observations were largely non-participant in a seat at the back or the side of the room, however, there were some interactions with teacher and pupils when either invited me to take a closer look or join the activity. Observation both 'disturbs', 'shapes' and is 'shaped by' what is observed (Lincoln and Guba (1985: 98). The presence of an observer would affect the lesson whether non-participant or not, with perhaps both teacher and pupil attempting to present the 'ideal' lesson since I had been introduced as a visitor looking at science, however, the aim was not really to judge the typicality of practice, it was more about trying to describe what is possible in assessment.

Observation is more than just looking, the gathering of 'live' data can be both systematic and selective (Cohen et al. 2011). The decision was made not to use recording equipment since this would be much more intrusive and would accentuate the feeling of 'being watched' rather than building a quality relationship between the 'knower and the known' (Lincoln and Guba 1985) in the collaborative nature of research promoted by the DBR approach. Field notes are necessarily selective, with only part of the lesson being recorded. However, in order to provide a more complete picture, field notes were supplemented by lesson documentation (planning and pupil work) together with pre and post lesson

discussions with the teacher. The way teachers used assessment was the key line of enquiry, thus there was a focus on the practice of the teacher during each lesson observation.

Open field notes were taken during the lesson observations. This was supported by an observation schedule completed after the lesson (Appendix 3D) which contained assessment features from Harlen (2013). The assessment features were included in the observation schedule in order to 'test' the categories in practice, as part of the DBR process. Many of the Harlen (2013) assessment features became the basis for the details within TAPS pyramid layers (Davies et al. 2014). The observation began descriptively, then became more focused and selective (Flick 2009), by organising the notes at the end of the lesson with the developing TAPS pyramid categories, for example, teacher elicitation, observation or discussion (Harlen 2013). Making descriptive notes on the lesson events was not a completely inductive approach because the TAPS pyramid categories were in mind, but by recording the chronology of the lesson before categorising, I hoped to reduce the bias or 'expectancy effect' inherent in looking for behaviour to match categories (Newby 2010, Cohen et al. 2011). This also meant that for each lesson there were two sources of observer data: chronological field notes and lesson events mapped onto the TAPS categories. The observation schedule developed over the course of the project, as the TAPS pyramid evolved; for example, by March 2014 the pyramid had an additional pupil layer (Short 2014), and so these categories were added to the observation schedule (Appendix 3E).

3.4.5 Discussions and interviews within case studies

A key method for exploring the relationship between formative and summative assessment was talking to participants. Discussions with teachers took place on development days and on school visits, ranging from a brief pre or post lesson conversation to a more formal semi-structured interview. Where possible the interviews were audio recorded, but the majority of discussions were informal and in public places, for example, in the staff room or at the reception desk in the case of one head teacher, and so field notes were taken during many of the discussions. For all interactions it was essential to build a rapport with the participants, a trust that would be conducive to open discussion and sharing of ideas, experiences and thoughts (Mears 2012). This required open listening, but also a recognition that these interactions are social encounters, the 'inter-view' as an interchange of views

(Kvale and Brinkmann 2009). Such social encounters involve interaction on both sides, thus I found that the conversation flowed more easily when there was not a strict script to follow or an audiotape being used. Semi-structured interviews are designed to be flexible, with key open-ended questions aimed at supporting the participant to speak freely (Cohen et al. 2011) and additional follow up questions planned but not necessarily used. This was fit for my purpose because I was looking to stimulate in-depth responses, rich data to understand a case (Newby 2010), rather than comparison of answers for particular questions across the population (Oppenheim 1992).

The informal discussions were related to issues arising from the lesson observation or other activity. Abrahams and Millar (2008) suggest that a combination of observation and interview enhances ecological validity because the interviewee is likely to be more anchored to reality rather than 'rhetoric' if the interviewer has observed the practice. This perhaps suggests that there is likely to be a closer match between espoused and enacted practice, but it also may be that in conceptualising their practice, the teachers are able to draw upon a shared experience of an observed lesson to exemplify and explain their approach to the interviewer.

The questions for the semi-structured interviews were compiled from the developing theories as part of the DBR process: initially using the layers from the Nuffield model (2012) for the first school visits in November 2013 (Appendix 3F) and later using elements from the developing TAPS pyramid. The final SL interview for School B explored experiences of the TAPS project (Appendix 3G) and was carried out by a different researcher so that the SL could speak to someone other than their link tutor, with whom they had been closely involved for three years. This meant that the SL needed to explain their assessment approaches afresh to a new person, this both guarded against SL concerns about repeating themselves, since data given to their link tutor may have reached 'saturation point' (Angrosino 2012), and could provide a point of triangulation as they explained from scratch.

A large amount of data, using a range of methods was collected, an essential ingredient of a case study as a basis for exploration of the significant features of the case (Bassey 1999). Whilst all the data helped to create a rich picture of each case, not all of it was directly

relevant to this study of the relationship between formative and summative assessment. Therefore, during the data analysis phase of the research, coding focused on the RQs, with some of the later lesson observations for example, receiving less attention because at that time the TAPS project focus had moved onto the development of assessment activities.

3.5 Validity and reliability in social science research

The terms used to describe confidence in a piece of research are wide-ranging and a source of debate. Many qualitative researchers reject positivist notions of validity and reliability, suggesting that they emphasise a search for a 'truth' which does not recognise that in a socially constructed reality there will not be one 'truth' (Sprague 2010). In this study, the terms 'validity' and 'reliability' also have a particular resonance for the topic of the research, and there are similar debates regarding the replicability of an assessment and whether it can uncover some underlying intelligence contained within the child. For both research and assessment purposes, I will explore validity and reliability, defining and utilising the terms as concepts from an interpretivist viewpoint.

3.5.1 Validity and trustworthiness

The core idea of validity is that the interpretation of data in particular ways should be 'explicitly justified' (Coe 2012: 42), thus validity is a concern throughout the research process. Threats to validity cannot be completely removed, but attention to validity at each stage of research can minimise their effects (Cohen et al. 2011). Coe (2012) argues that validity is a fundamental but 'confused' concept, because it is often focused on the data collection instruments, rather than the interpretation. Researchers may focus on methods, assuming that by performing certain 'checks' validity would be guaranteed, but: *"validity threats are made implausible by evidence, not methods"* (Maxwell 2010: 279).

Interpretivists may suggest the term 'validity' is not relevant for consideration of social actions (Bassegy 1999), instead Lincoln and Guba (1985) propose 'trustworthiness', moving from a search for 'truth' to a judgement of the persuasiveness of the findings. The audience for the research evaluate its trustworthiness and decide whether they are sufficiently

persuaded for the findings to be utilised and acted upon; validation is reformulated as the social construction of knowledge (Mishler 2010). However, others argue that the term validity does not necessarily imply the existence of an 'objective truth'; validity is relative to the specific purpose and circumstance of each piece of research (Maxwell 2010).

3.5.2 Internal validity and credibility

Internal validity or credibility (Lincoln and Guba 1985) is concerned with the way the collection of data leads to the conclusions; the extent to which the researcher's explanations and interpretations are supported and sustained by the data (Cohen et al. 2011). As noted above, validity is a concern throughout the research process, so a number of sections discuss how the credibility of the research has been enhanced. The combination of case study, with its in-depth rich data and thick description (Geertz 1973), was complemented by the broader analysis of a wider dataset, providing different lenses through which to examine practice (Section 3.3). Transparency in the processes of data collection (Section 3.4) and analysis (Section 3.7) provides for full and explicit audit trails so that the reader can see the basis for plausible interpretations (Coe 2012). This study was made more rigorous through the use of methods, time, investigator and source triangulation (Section 3.5.6), together with respondent validation (Section 3.5.7). In addition, the way the PhD study was placed within a larger research project, provided further opportunities to enhance validity. Long term access to schools and data collection provided prolonged engagement (Lincoln and Guba 1985) and being part of a wider research team provided the opportunity for methods, findings and interpretations to be discussed and 'tested' throughout the process.

3.5.3 External validity and transferability

The question of external validity of research was introduced with regard to case studies in Section 3.3.3 on 'fuzzy generalisations' (Bassegy 1999). Lincoln and Guba (1985) suggest 'transferability' is a preferable term because it places the focus on the researcher providing rich description, from which working hypotheses can be drawn about that context, making no assumptions about other contexts in the population. Greene (2010) notes that the notion of transferability shifts the judgement from the inquirer to the potential user as they

judge the applicability of the findings to their context (p69), similar to Stake's (2006) description of 'natural generalisation'. A criticism of this approach is that it is unclear how rich the description needs to be, in order to enable the potential user to generalise to their own context. This would appear to be very much about the audience to which the description is aimed, for example, for a classroom teacher the age of the pupils in the study may be an integral piece of information, whilst a researcher may be more interested in school size or locality. A more striking point for this view of external validity as transferability, is that the: *"interpretivist inquirer must provide for the possibility of transferability, but not its actualisation"* (Greene 2010: 70). The suggestion that a criterion of external validity is satisfied if transferability is a possibility, whilst it makes research more manageable, also appears to miss the point of research having an impact, a requirement embodied in the concept of consequential validity.

3.5.4 Consequential validity and authenticity

Stake (2006) suggests that it is the author's responsibility to support the readers' interpretations by repeating key assertions in several ways, with illustrative examples, in order to respond to Messick's (1989) criterion of consequential validity (cited in Stake 2006: 35). Consequential validity is transferability in action, where the research stimulates thinking, understanding or action; this has also been called catalytic validity (Cohen et al. 2011). If the research is used as the basis for further research or changes in practice, then it marks an overall assessment of trustworthiness (Mishler 2010). A feature of authenticity in research is to empower action (Simons 2009) and this is very much in line with the aims of Design-Based Research. Of course, such consequential action following research, may not always lead to desirable outcomes, thus it is the responsibility of the researcher to consider the most likely uses or interpretations of the research (Coe 2012). Transparency in research processes, explaining how the data supports the validity of the conclusions, enables it to be both subjected to public scrutiny, but also enables future research to be based on the work (Hedges 2012: 23). DBR aims to create theoretical and practical products which will support practice, so this study must result in such products if it is to be consequentially valid.

3.5.5 Reliability and dependability

Reliability is concerned with the accuracy and consistency of findings. For positivists, replicability is key for external reliability, however, this is not the same for interpretivists since each situation is unique and socially constructed. In addition, each point of data collection, particularly if it involves face-to-face interviews or observations, is a context which includes the researcher. In DBR the process of research and its products are inextricably linked; the interaction between the researcher and the researched is an explicit part of knowledge production (Evans 2013). Reliability in qualitative research could be seen as the 'fit' between what the researcher records as data and what actually occurs in the setting: *"a degree of accuracy and comprehensiveness of coverage"* (Cohen et al. 2011: 202). Lincoln and Guba (1985) use the term 'dependability' for describing the rigour of the research. Dependability is strengthened in similar ways to internal validity, with for example, respondent validation and triangulation which will be discussed in the next sections.

Quantitative researchers would consider an important feature of internal reliability to be inter-rater reliability. For qualitative researchers, it is useful to consider the viewpoints of different researchers, but this may be more to gain insights into a range of possible interpretations, rather than explicit checking the recording of a 'true reality'. Qualitative tools are not neutral and standardised (Mason 1996), they aim to capture the complexity of social situations. Nevertheless, it is important that researchers demonstrate the rigour of their research with transparent and detailed descriptions of how the research was carried out, together with explicit explanations of how conclusions were reached, including where theory supported particular interpretations (Silverman 2011). Co-researchers need to agree on the meaning of key terms for their research, for example, when recording and categorising observations, where this study overlaps with the larger TAPS project. A transparent process for coding was important in this research (see Section 3.7), in order to consider the consistency across the dataset. There is no suggestion here that the same event is repeating over time, but that the researcher looks for and defines codes or themes which are repeatedly present in the range of data available. This consistency across time and instruments, combats accusations of anecdotalism (Silverman 2011), supporting the reliability or dependability in qualitative research.

3.5.6 Triangulation

A widely used strategy for strengthening the internal validity of research is triangulation, which aims to reduce the effects or bias inherent in particular data collection methods (Evans 2013). Early conceptualisations of triangulation likened it to a geographical pinpointing, navigating to a certain point by cross-referencing from different points of data. This view of triangulation is more in line with the positivist tradition, with multiple sources of data converging to pinpoint the 'truth' of the reality (Mason 1996). Under this view, the context and social interaction is 'ignored' as the researcher repeats observations to try to uncover the 'truth' (Silverman 2011). However, triangulation has a different emphasis for interpretivist researchers; it is concerned with developing a richer, holistic picture of the context through exploration of multiple perspectives, together with how they are constructed and related (Simons 2009). It is less about the repeatability of an observation, but more about clarifying the meaning by identifying the different ways the case is being seen (Stake 2006).

In this study, triangulation has been used to strengthen the case study research in the following ways (Cohen et al. 2011):

- Methods triangulation: using a range of methods, for example, observation, interview, written tasks. The same methods were also used in different contexts or on different occasions, for example, observations in different classes. (Section 3.4).
- Time triangulation: ongoing involvement and data collection with each school for a two year (School A) or three year (School B) period.
- Investigator triangulation: on three occasions different researchers collected data (first school visits and Subject Leader interview for School B July 2016).
- Source triangulation: seeking to involve a number of teachers from each school (discussed below).

The PSQM data and much of the case study data originated from the school's science subject leader (SL), information from those in a position to give it (Teddlie and Yu 2007). The SL is the external facing school representative for science, together with leading, monitoring

and developing science within the school. In effect they act as gatekeeper for science, and it is through them that data about the school can be collected. As the person responsible for coordinating science in the school, it is likely that the SL is more confident than other staff in the teaching and assessment of science, thus the practice in their classes may be different from other classes in the school. In order to work with the SLs over a long period of time, it was important to build a rapport with them which would facilitate access to the classrooms, documents and viewpoints held within the school. Where possible, for triangulation within the case studies, viewpoints other than those of the SL were sought. It cannot be assumed that views are uniform across the school, however, for simplicity, the case studies will refer to School A and B when exploring the different perspectives to develop a deeper understanding of the relationship between formative and summative assessment.

Triangulation is not something only to be considered at the point of data collection; it is an essential ingredient during data analysis (see Section 3.7), with each interpretation both checked that it is supported by existing data (Stake 2006), together with comparison to other data and theory (Coe 2012).

3.5.7 Respondent validation or member checking

Another way to strengthen the research is to verify data, findings and interpretations with participants. In this study, this took the form of: checking field notes and other data for accuracy with the teachers after the school visits; discussing interpretations with teachers at development days and over email; providing participant access to stored data; and sharing draft case reports with subject leaders. Such a list makes the process sound like a simple checking of wording, however, respondent validation is more than this; it is concerned with fairness of the portrayal and has ethical implications for the participants (which will be discussed more fully in Section 3.6). Silverman (2011) argues that respondent validation is problematic if the data, or its interpretation, are not compatible with the participant's self-image; such reflections could provide a further source of data, but also raise questions about whether the data should be used if it is not comfortable for the participants. Newby (2010) suggests that when considering evidence where participants may represent

themselves and their circumstances in a 'better light', for example, describing formative assessment which is not formative in practice, triangulation of data is useful.

In this study, it was important for the participants to be active in the research process, in line with the principle of collaboration and co-research in the Design-Based Research approach. Together with this there is a relationship building which takes place between researcher and participants in longitudinal case study, which is integral to gaining access and 'quality' data (Simons 2009). The close relationships built with the teachers in the case study schools, together with the promotion of their active role in the research, made respondent validation an integral part of the process. However, it also meant that the level of critical analysis of individual incidents and perspectives needed to show an awareness of the effect this may have when read by a participant; thus some of the critique is more fully explored in Chapter 7, rather than the case study chapters.

3.6 Ethics

3.6.1 Key principles

Proponents of situated ethics question whether it is possible to have universal abstract rules since ethical principles take on different significances in different research practices and ethical decision making should take place throughout the research process (Simons and Usher 2000). Thus ethical principles are discussed as guidelines rather than rules (BERA 2011) since there is much decision-making required in their application.

In this study, the two types of sample required different approaches to consent, since the PSQM sample was a pre-existing dataset, whilst the case studies were ongoing and evolving, as will be discussed below. The PSQM Round 4 subject leaders had written assessment reflections for the purpose of gaining an award, not for the purpose of providing data for a study. However, at the point of submission the PSQM website stated that uploads may be used for research purposes. Nevertheless, I felt that although the participants had been aware that their data may have been used for research, they had not had the right to withdraw (BERA 2011). Therefore, the 91 subject leaders were contacted by PSQM (so that

their anonymity was preserved) to inform them of my study, how results would be shared and given an opportunity to remove their data from the study. One replied to ask for a copy of the article (Earle 2014), but no schools requested that their anonymous data be removed.

The ethical principle of respect was particularly important for the case studies, the face-to-face contact with participants required a responsibility to treat each individual fairly and 'to do no harm' (Luttrell 2010). In my roles as classroom teacher and teacher trainer, I had gained an understanding of such professional duty of care, together with experience regarding the need for an approach which was sensitive to each school's climate and each individual's context.

Voluntary informed consent pertains to both the nature and implications of the research for participants (Homan 2002). The teachers from each school signed a consent form at the outset of the TAPS project (Appendix 3H), however, consent is an ongoing process (Luttrell 2010) and there were many other times when consent was sought during the research project: to observe lessons, to audio record interviews, to utilise PSQM submissions within the PhD research and to feedback on case study drafts. In order to minimise a 'duty-bound' response to being involved in the PhD research, consent was discussed with participants who played an active part: *"as thinkers about their lives rather than data producers"* (Luttrell 2010: 4). Consent discussions throughout and after the research were largely verbal or via email, since formalising the process each time in a written form would re-assert a power relationship not commensurate with collaborative research (Usher 2000).

The PSQM Round 4 data was anonymised at the point of download from the website, with each school being given a numerical code. The case study data was anonymised at the point of writing, for it was accepted that: *"anonymity as a protective device means little since it cannot be applied to the face to face encounters with the researcher"* (Ulichny and Schoener 2010). Secure storage of data, including the raw case study data which had not been anonymised, was achieved by password protected university google folders and locked storage for paper data. Whilst privacy was protected for academic writing, there was the option for TAPS schools to share practice in reports or online which included the name of their school or science subject leader. There is a clear divide here between the named

sharing of good practice and the critical analysis of anonymous practice. Nevertheless, it could be possible for an interested individual to cross-match examples from academic writing and public examples to identify the participants in the former. This begs the question as to whether it is possible to ever fully anonymise studies of social practice (Mason 1996). I endeavoured to ensure participants were able to express opinion and choice regarding their data, for example, providing the participants with drafts of academic writing enabled them to comment on and improve the explanation in case study chapters.

3.6.2 Roles of researcher

Ethical guidelines may assume a clear and unchanging delineation between researcher and researched but ethical concerns are not static, particularly in this dynamic study (Haney and Lykes 2010: 112). Both the study and the relationships within it evolved: I began as a link tutor, not long out of the classroom, but still in a position of power as a university researcher who was studying school practice. The more visits and meetings that took place, the more I came to know the teachers and the more we approached the discussions collaboratively. But as I took on the role of TAPS project lead and spoke at more TAPS events, perhaps the balance of power in relationships with teachers tipped again as I returned back to the role of 'expert'. A certain distancing is essential at the analysis stage of case study research as I sought to explore the data from different viewpoints, so in the second phase of TAPS I appointed different link tutors to the case study schools. Nevertheless, my relationship with the subject leaders continues as I invite them to explore my interpretations of their work.

Ulichny and Schoener (2010) describe two aspects of the research role: the action taken by the researcher, ranging from distanced observer to full participant; and the relationship between researcher and subject, ranging from stranger to mutual acknowledgement to friend (p422). My role was multi-faceted so I can identify times where I was at each end of the spectrum, for example, I would sometimes observe at a distance for part of the lesson, but at other times we would work together to try to pick out key parts of the lesson, actively trialling innovation (Kelly 2003). I found that to maintain a silent observation was unnerving for the participants, for example, when I observed a staff meeting silent note-taking

appeared to make the teachers nervous, thus there needed to be a balance between observation and interaction. To a certain extent, the participants needed to see me as part of the team, or 'on their side' to be able to accept me into their classroom and confidence. In addition, it was important to find time to discuss observed lessons with the teachers, a lack of feedback could be interpreted as 'silent criticism' and lead to a lack of trust (Ulichny and Schoener 2010: 426).

3.6.3 Roles of participants

The case study participants were class teachers from TAPS project schools, some of whom also held leadership responsibility for science or the school. By being part of the TAPS project, the teachers perhaps felt part of a team working towards a common goal and so were more open than they would have been to a single researcher. Since the whole school were part of the TAPS project then individual teachers did not feel singled out; an institution-wide investigation can feel less invasive since attention is on the whole context rather than on an individual's practice (Ulichny and Schoener 2010: 425).

The children in the schools form part of the context, and ultimately it is their learning that the project hopes to enhance, but the focus for both TAPS and the PhD is on teachers. Anonymised children's work forms a small part of the data, as examples of assessments which were discussed with teachers, but this work was part of their everyday school work, rather than something done specifically for research purposes. The head teacher and class teachers were designated as the 'gatekeepers' to the work and it was their permission which was sought, as has been done in other studies which used children's everyday classwork (Haney and Lykes 2010: 115). Both schools had also informed parents that researchers would be visiting the school, and in each lesson observation, I was introduced to the pupils as someone who had come to learn about science at their school.

Ethically we can go beyond the principle of 'do no harm' and consider ways that the research can 'give back' to the participants (Luttrell 2010). There was a significant time commitment for the teachers joining this study and so it is important that they benefit from the process: "*not just academic research on others as subjects but rather inquiry with others*

to improve practice” (Haney and Lykes 2010: 112). This is also in line with Design-Based Research where research collaborations lead to new products and understandings. Thus both research design and a consideration of ethics suggest that participants should be more than subjects of observation, they should be full partners in the research, exploring and developing their practice throughout the project, rather than waiting for a report at the end. This makes the roles of researcher and participant more complex and dynamic than traditional research, but by considering process as well as product, the research can provide the ‘thick description’ (Geertz 1976) which the complexities of education demand.

3.7 Data analysis

3.7.1 Approach to data analysis

Qualitative data collection, particularly for the case study section of this research, resulted in a wide range of material, which supports triangulation (Section 3.5.6), but for which there are no clear cut rules regarding interpretivist analysis, to organise and make sense of data. Thematic or content analysis is widely used but could be described as more of a ‘cluster of techniques’ rather than an ‘identifiable approach’ (Bryman 2016: 584). Data analysis procedures often include breaking it down into segments which can be categorised and examined for connections, patterns and propositions that seek to explain the data (Simons 2009). Such procedures for organising the data need to be systematic and rigorous, with comprehensive examination of all avenues; the qualitative researcher cannot just respond to the ‘loudest bangs’ or ‘brightest lights’ (Cohen et al. 2011: 202). Transparency of these research processes is essential in order to justify how the validity of the conclusions is supported by the empirical evidence (Hedges 2012: 23). This transparency refers to both the research processes and the theoretical underpinning, thus it is important to make the theoretical stance explicit (Silverman 2011). Hence the rest of this section will explicitly describe the way content analysis was applied to the three datasets in this research.

3.7.2 Theory-led and emergent coding

Many qualitative researchers aim to take an emergent or ‘grounded theory’ approach, whereby analysis begins with the data and theory arises from it: theory ‘comes last’ (Mason 1996). However, in this research there was a significant amount of theory that informed the

project from the outset, with the Nuffield (2012) pyramid model of ‘formative to summative’ and the TAPS pyramid school self-evaluation tool (Davies et al. 2014). Thus a theory-led approach is visible in all stages of the research: in the construction of the RQs, in the data collection (e.g. lesson observation proforma and interview questions), in the pre-codes (list of theory-led terms to code in the data) and in the presentation of case study data analysis in TAPS pyramid layers.

Nevertheless, a purely theory-led approach could be open to criticisms of circularity and confirmation bias, where data is selected to fit the theory (Maxwell 2010), and research is isolated rather than cumulative (Hartas 2010). In addition, DBR aims to raise the participant to the role of co-researcher, and so it is important to give the participant a ‘voice’ in the research, through their data providing new lines of enquiry. Thus, whilst coding of the data began with a list of theory-led codes generated from the RQs and theory (Nuffield 2012, Davies et al. 2014), further emergent codes were also added as they arose from the data, as detailed in Table 3.3 and further explored below.

Table 3.3 Theory-led and emergent codes

First source of codes	Codes
Theory-led codes from RQs	formative, summative, purpose
Emergent codes from initial analysis of PSQM <i>(for full list of PSQM 2nd level analysis sub-codes and themes see Appendices 4B, C and D)</i>	APP (assessing pupil progress), tracking, tests, summative named by teacher, summative other, moderation, next step identified as moderation, next step for testing, AfL /formative named by teacher, elicitation strategies, learning objective, marking, gaps in learning, self-assessment, peer-assessment
Theory-led codes from TAPS pyramid layers analytical framework	act on feedback, planning, questioning/discussion, observation, recording/evidence, adapting, teacher feedback range of activity, shared understanding, criteria, recordkeeping, summarise, reports
Emergent codes from School A	strategies, confidence, consistency, portfolio, PSQM, TAPS, Subject Leader role, structures
Emergent codes from School B	challenges, differentiation, evidence, knowledge, levelling, marking, parents, self-evaluation, sharing practice, skills, staff team

3.7.3 The process of coding

The three datasets (PSQM, School A and School B) were analysed separately, but for each a similar qualitative content analysis approach was taken (Silverman 2011). Data analysis involved a ‘to and fro’ between the raw data and interpretations of it, a process of ‘constant comparison’ (Robson 2011). A brief description of the cycles of analysis is detailed in Table 3.4 below. The table layout suggests a generally linear pathway, but in reality there was an ongoing ‘to and fro’ between the data and its organisation which will be explored below.

Table 3.4 Overview of qualitative data analysis

	PSQM data	School A	School B
Initial organisation	SL reflections anonymised (numbered in download order)	Case record collated and anonymised. Chronological order to begin to break into DBR phases. (Appendix 5A and 6A)	
Initial reading	To understand the kind of information provided	To gain holistic sense of case record and consider where DBR phases might start/end.	
List of pre- codes (theory-led codes)	Listing of potential codes from the RQs	Listing of potential codes from the RQs and TAPS pyramid (Appendix 5B)	
Initial coding , including addition of emergent codes	Addition of codes arising from data (Appendix 4B)	Addition of codes arising from data, e.g. when repetition was noted (Appendix 5B)	
Revisit DBR phases	----	Align timing of phases with emerging picture from data	
Repeatedly re-read, pruning/amending codes	Formative/summative codes too broad, led to finer grained sub-codes for 2nd level of analysis	Checking codes are not too broad/narrow, removing or rephrasing codes (process repeated until clarity and representativeness in coding)	
Cross-checking internal consistency of codes	Checking coded material is consistent with other material of the same code		
Second level of analysis: Identification of ‘higher order codes’/themes	Sorting into detailed categories (Appendices 4C and 4D)	Return to RQs to identify ‘higher order’ codes which are recurring in the data and relevant to the RQs	
Consider frequency of codes for different DBR phases	----	----	Consider how frequency changes (App 6B)
Initial selection of quotes	Quotations selected which were representative of code/theme		

Consider different organisations of 'higher order codes'/themes	----	Map emergent themes onto RQs, DBR phases, TAPS pyramid layers, and principles of validity/reliability/manageability.	
Draft organisation of 'higher order codes'/ themes	Codes grouped by RQ: formative and summative	Codes/themes grouped by RQs	Codes/themes grouped by emergent themes.
External validation	Feedback on draft themes and interpretation from JT	Feedback on draft themes and interpretation from DD	Feedback on draft themes and interpretation from KM
Final grouping of 'higher order codes'/ themes	Formative and summative themes further explored e.g. role of self-assessment.	Codes/themes grouped by pyramid layers (App 5C)	Codes grouped into emergent themes by DBR phase and pyramid layer (Appendix 6C)
Final write up	Including re-visiting data, codes and themes		

ATLAS.ti software was used to support the data analysis, as an administrative tool for coding and retrieval of data, rather than a tool for constructing themes since the largely deductive theory-led approach was less supported by software which had been developed with 'grounded theory' in mind (Gibbs 2012). However, the 'code and retrieve' use of the software enabled a large number of sources to be examined and efficient 'constant comparison' as coded items could easily be retrieved and compared, supporting consistency checks and refinement of coding definitions. 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. A point of 'theoretical saturation' was reached when new codes are no longer 'illuminating' (Bryman 2016: 573).

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, proposing next steps rather than listing current 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 for two purposes: to support analysis of the prevalence of an assessment strategy (for example, in the PSQM data), and to support checking of the analysis to 'combat anecdotalism' (Bryman 2012: 624); so that a memorable event had not taken undue precedence in the data (for example, in comparison of the DBR Phases for School B).

The process of coding qualitative data, and computer-assisted data analysis in particular, has been criticised for distancing the researcher from the data (Gibbs 2012), and for a loss of context (Bryman 2016). As a DBR researcher, working closely with the participant schools, the sense of distance actually proved to be useful, enabling me to look at the data afresh. Also, the 'constant comparison' over time, provided rigour to the analysis, with codes being cross-referenced and checked with the data sources, with newly collected data compared with existing data, enabling both theory and coding to be revised (Coe 2012).

3.7.4 From codes to themes

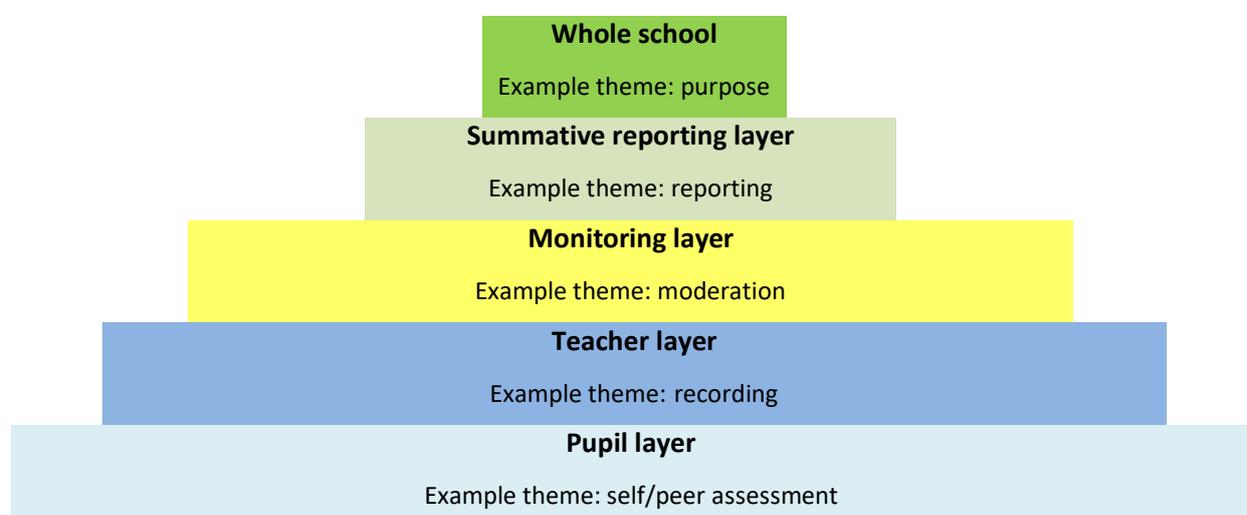
Whilst the coding of individual items had decontextualized and fragmented the data (Bryman 2016), the process of grouping and constructing themes is designed to bring the data back together again. Having spent so much time in the 'to and fro' of coding, a number of the codes had emerged as 'higher order codes' in the case study data, hence these codes became the themes for discussion, and are distinguished in bold in the data analysis chapters. By 'higher order codes' I mean that they were both recurring codes and pertinent to the RQs, for it needs to be more than repetition that designates a theme, it must relate to the RQs (Bryman 2016: 586).

A number of the codes which had emerged from the data were not directly relevant to the research questions (for example, 'parents'); these codes appear in the appendices, but are not a focus for discussion in this study. A full list of codes for each dataset can be found in Appendices 4B, 4C, 4D, 5B and 6B.

A number of analytical frameworks were explored to structure the case study data analysis: my RQs, DBR phases, the TAPS pyramid and the assessment principles of validity, reliability and manageability. These analytical frameworks provided different lenses through which to examine the data and my interpretations, challenging claims of circularity for selection of ‘higher order’ codes. By looking at the data from many angles, and repeatedly revisiting the data, an in-depth understanding was gained from a range of viewpoints. In response to external feedback and in an attempt to represent comprehensive coverage of the data from across the school, the case study chapters were organised by TAPS pyramid layer (Davies et al. 2014). This also provided the opportunity to ‘test’ the ‘formative to summative’ model in action, exploring the relationship between formative and summative assessment, which had been identified as lacking clarity in the model (Davies et al. 2017). Relevance to the RQs was the key to the selection of a ‘higher order’ code, whilst the TAPS pyramid layer designated where in the thesis the code was discussed. Figure 3.1 provides an outline of this structure and an example ‘higher order’ code or theme which was representative of each layer. School B data was additionally grouped by DBR Phase to allow for comparison across time. A full mapping of case study codes to TAPS pyramid layers can be found in Appendices 5C and 6C, together with a ‘higher order’ codes mapping at the start of each case study chapter.

Figure 3.1

TAPS pyramid analytical framework: pyramid layers and themes (‘higher order’ codes)



3.8 Summary

The key methodological decisions discussed in this chapter were:

- The aim to develop theory and practical guidance led to the selection of an interpretivist Design-Based Research methodology in which collaboration with practitioners is utilised to meet dual goals of developing both theory and practice through iterative cycles in real contexts.
- To answer the RQs regarding the conceptualisation and enactment of the relationship between formative and summative assessment, two types of sample were selected: a pre-existing PSQM dataset which contained descriptions of assessment practice from schools across England; and case studies of assessment practice in action over time. The two cases selected were purposive: ‘theory-seeking’ and ‘theory-testing’ (Basse 1999).
- Data collection methods were selected: documentation; non-participant observation; semi-structured discussions/meetings; and written tasks.
- Validity and reliability of the research were enhanced by: transparency in the processes of data collection and analysis; triangulation of methods over time; and respondent validation.
- Ethical principles were considered throughout the study, including ongoing discussions with participants regarding consent during the developing and iterative cycles.
- Qualitative content analysis included both theory-led and emergent coding, with ‘higher order’ codes pertinent to the RQs becoming themes for discussion and placed within the TAPS pyramid layers framework (Davies et al. 2014).

This chapter introduced Design Based Research methodology which aims to develop both theory and practice, together with the range of qualitative methods used to answer the research questions in this study. The layers from the TAPS pyramid school self-evaluation tool (Davies et al. 2014) provided a way of structuring the case study data (Figure 3.1), but this was not yet in existence at the time of the PSQM data analysis, so the next chapter provides a preliminary study, organised to provide a mapping of current practice in formative and summative assessment in order to answer RQ1: how do teachers assess children’s learning in science for formative and summative purposes?

Chapter 4 Formative and summative assessment in submissions to the Primary Science Quality Mark

4.1 Introduction

The literature review identified a lack of clear UK government guidance for formative and summative assessment in primary science in a climate of assessment reform (DfE 2013b). This chapter seeks to address Research Question 1 (RQ1) regarding current practice in formative and summative assessment in primary science:

***RQ1.** How do teachers assess children’s learning in science for formative and summative purposes?*

In order to gain an understanding of the landscape of primary science assessment, a pre-existing dataset from the Primary Science Quality Mark (PSQM) was used to map the approaches taken by 91 English primary schools (at March 2013). PSQM requires the science subject leader (SL) in each school to reflect upon and develop practice over the course of one year (in this case March 2012 - March 2013), then upload a set of reflections and supporting evidence to the database to support their application. One of the criteria (C2) requires the SL to explain in writing how science is assessed within the school and it is these reflections which were anonymously analysed for all completed Round 4 PSQM submissions (see Appendix 4A for details of the data and a secure link to a folder containing the source material). As discussed in Chapter 3, qualitative content analysis, supported by ATLAS.ti software, was used to code descriptions of formative and summative assessment (for a full list of codes and their organisation, see Appendix 4B, 4C and 4D). The data presented in this chapter includes numerical counts, to show the prevalence of assessment strategies in the sample, together with examples of the coded extracts, in order to discuss the use of the strategies, as well as supporting transparency of the analysis.

The chapter will begin with an exploration of the methods used for summative assessment described in the dataset (Section 4.2), including a mapping of the range and combination of methods present in the sample. Section 4.3 will focus on the strategies used for formative assessment described by the Round 4 schools, including the different ways schools elicited pupil ideas and how schools described pupil self-assessment. Section 4.4 will consider the relationship between formative and summative assessment presented in the sample, although the lack of explicit explanation in this area provides a clear focus for the case studies in the next chapters.

4.2 Summative assessment

4.2.1 Categorising as summative

The PSQM C2 reflective paragraphs contained descriptions of school assessment. In order to classify the sections which concerned summative assessment, simple key word frequencies were not suitable, since subject leaders discussed the merits of different strategies. 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. 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'.

This section will present a range of examples to exemplify the descriptions of summative assessment. Section 4.2.2 will explore the frequency of such strategies within the dataset.

Most teachers wrote about summative assessment in their school, with only 2 of the 91 schools containing no mention of it. Some described a single method for summative assessment, for example, in this school they used Assessing Pupil Progress (APP), a particular type of tracking grid (which will be discussed further in Section 4.2.3):

Extract 4.1

In order to track children's attainment, it has been decided that all class teachers will be responsible for completing the APP for SC1 at least once every half term.

PSQM C2 reflection, March 2013 - **R4.14.5**

Coded as: APP tracking grid

In Extract 4.1 the method is named (APP), but the purpose and process is less clear since there is no explanation of how teachers complete the APP grid each half term. The subject leader describes the purpose of this as tracking of pupil attainment, but does not explain whether such tracking is primarily for accountability purposes or for identification of gaps in pupil attainment which could be used to support future learning (Mansell et al. 2009). This extract is typical of the PSQM C2 reflections, in that method or strategies are named but the processes are not fully explained.

My practitioner knowledge of the types of products used by schools was useful for understanding some of the PSQM reflections, with the subject leaders using names or abbreviations without explanation. For example, having recently been a teacher and continuing to work with teachers meant that I knew the 'Rising stars' product described in Extract 4.2 below was a scheme which included a set of tests. However, my experience as a teacher also meant that I needed to be aware of my assumptions when coding and categorising the submissions. Systematic checking of my interpretations was supported by the iterative cycles of data analysis which led to repeat readings of the SL reflections, each time returning to the PSQM submissions to ensure they were categorised fairly (Cohen et al. 2011).

Extract 4.2

Rising Stars is used to assess knowledge and understanding at the end of the topic...Year 6 also complete previous years SATs papers despite them no longer being statutory.

PSQM C2 reflection, March 2013 - **R4.15.3**

Coded as: Tests

Extract 4.2 includes a comment about use of non-statutory SATs papers, but the SL does not go on to explain what they think of this. It could be that they are trying to show that they are doing more summative assessment than is required, which could be a point about teacher workload or a point about checking because of uncertainty about other methods. Further data would be needed to explore the SL's viewpoint, but it is useful to note that some schools were using more than one type of test for summative assessment.

Many of the subject leader reflections described a combination of approaches, for example, in this case they used tests combined with teacher review of level descriptors:

Extract 4.3

We decided as a whole staff to use objectives from key skills books and to tick objectives which have been taught weekly and met and initial those who haven't quite met the objective or who has exceeded the objective taught. Furthermore, we have combined this with a more summative approach as test schemes are used at the end of a topic to support teachers with levelling and tracking progress.

PSQM C2 reflection, March 2013 - **R4.22.3**

Coded as: Tests to back up teacher judgement

This extract describes how lists of objectives were used to check for coverage of the curriculum and for recording the initials of children who had 'not met' or 'exceeded' weekly objectives. End of topic tests were used to 'support teachers with levelling and tracking progress'. Both methods appear to be summative, with the tests providing a level and the teachers making judgements against objectives (Taras 2005), but it is not clear how the information from ongoing weekly assessments were 'combined' with the information from tests. The PSQM dataset was sufficient for categorising assessment approaches, but did not provide a full explanation of the school's approach or processes, which necessitated the need for the case studies in Chapters 5 and 6.

4.2.2 Methods used by schools for summative assessment

In order to map the reported assessment techniques across the sample, each description of summative assessment was coded and tallied (see Appendix 4B, 4C and 4D for further details). The categorisation of summative assessment methods can be seen in detail in Figure 4.1 and in summary form in Figure 4.2. Analysis of statements from the 91 subject leaders found that only 2 did not explain how they assessed science summatively, thus the percentages in this section are based on 89 schools.

Figure 4.1 Summative assessment (detailed) for PSQM Round 4 (March 2013, N=91)

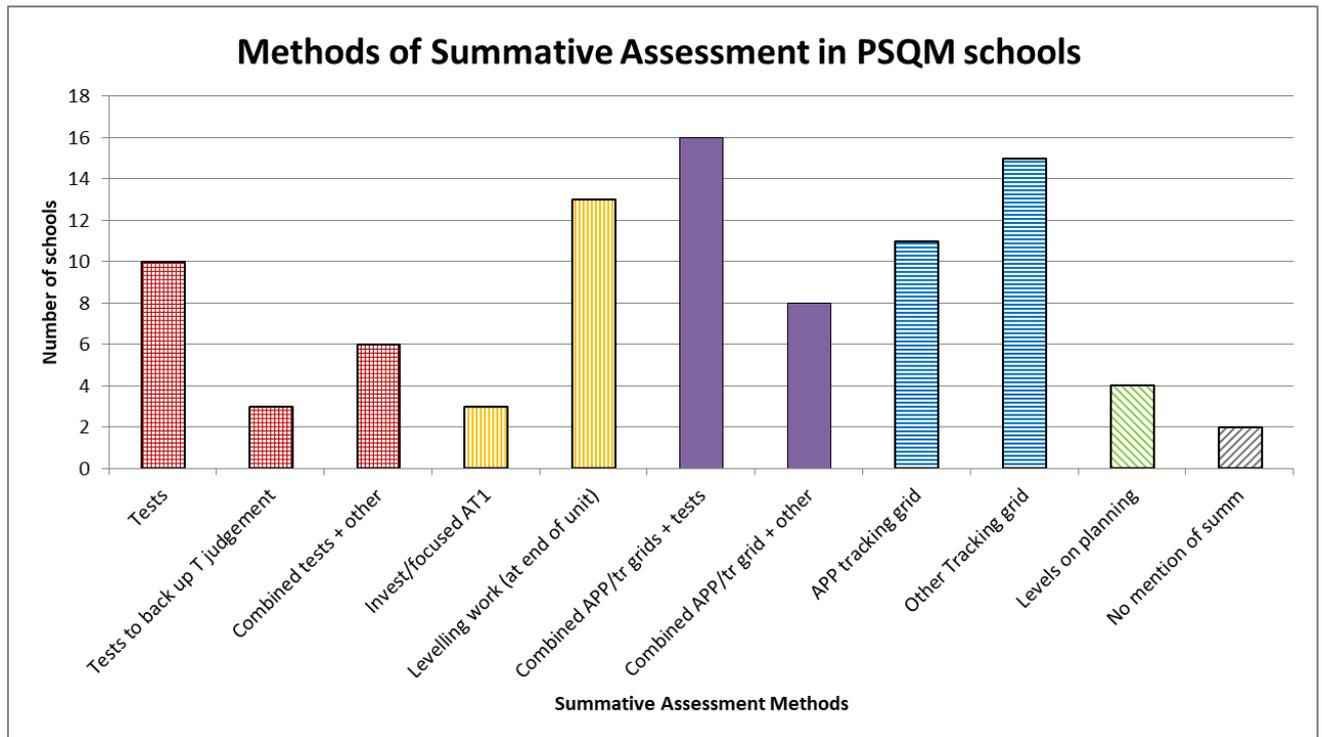
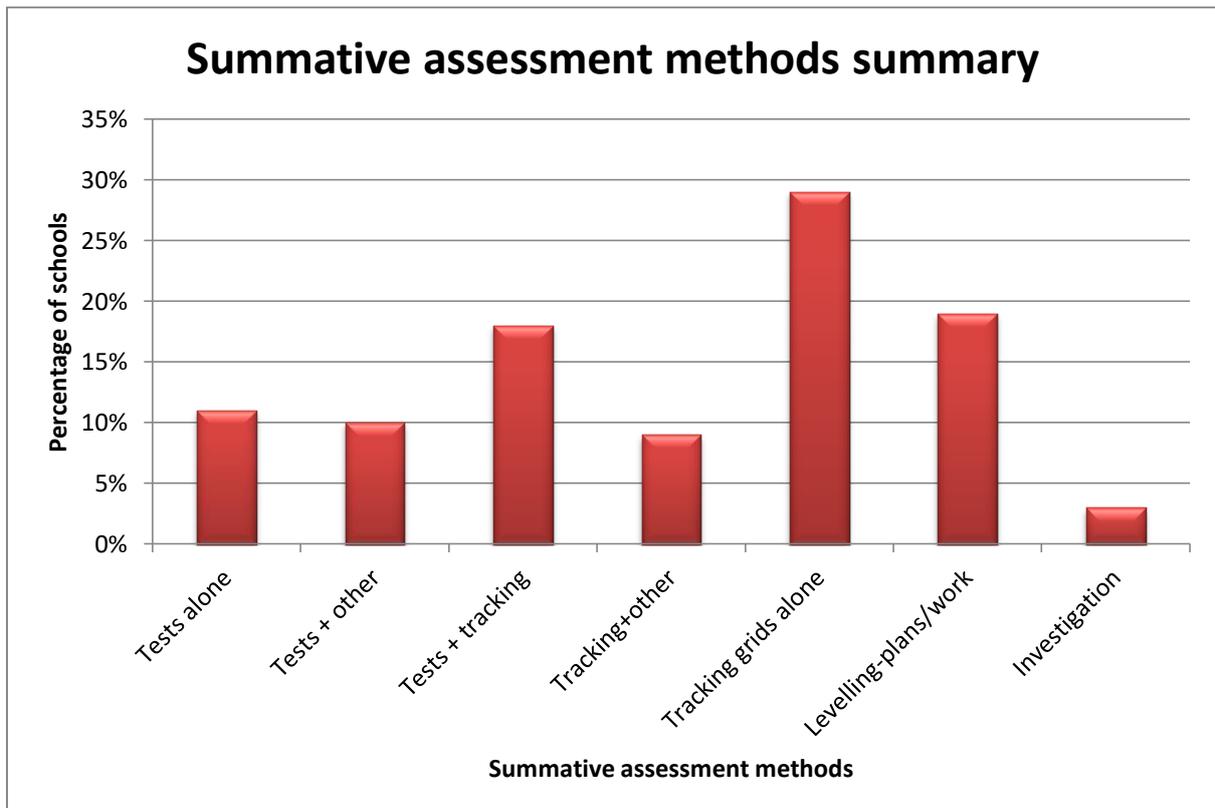


Figure 4.2 Summative assessment (summary) for PSQM Round 4 (March 2013, N=89)



NB. % based on 89 schools since 2 did not specify summative assessment methods.

Figure 4.1 and 4.2 demonstrate the range of summative assessment practice described by the sample, with the majority reporting use of a mix of methods. The use of tracking grids will be discussed further in Section 4.2.3, thus this section will focus discussion on the other major category of testing.

Many schools (38%) mentioned testing, but only 11% of this sample used testing alone (see Figure 4.2). The others used test results as part of the information, combining this information with other methods such as tracking grids. The distinction between using test information alone and using test information as part of the summative assessment is an important one because of concern over the validity of written tests to be able to sample the full range of objectives for primary science (Gardner et al. 2010), together with the reduced time available for practical work when test revision takes precedence. Schools in this sample had moved from compulsory SATs levelling in 2009 to only 11% using testing as their sole measure for summative assessment, albeit in schools where developments in primary science were a priority for the PSQM award. However, it is unclear from many of the reflections how the teachers are combining the information from different sources, for example:

Extract 4.4

'Rising Stars' is utilised at the end of each unit to assess pupils' knowledge and understanding of the topic. This method gives an overall level at the end of each topic and then at the end of the year. Since implementing this assessment tool throughout the school, teachers have reported that they are more confident in levelling children's work. Information from Rising Stars assessment is then correlated with our own assessment which we use for each of the topics. This is completed through observation, work and small group activities and evaluated at the end of each lesson.

PSQM C2 reflection, March 2013 - **R4.15.1**

In this extract, the subject leader described that the 'Rising Stars' tests, used to provide topic and end of year levels, was 'correlated' with teacher assessment information drawn from other lessons. This could mean that the information is combined and the teacher decides on an overall level, or it could mean that the tests are used for checking or standardising the teacher assessments. The teacher assessment activities are 'evaluated at

the end of each lesson', suggesting a possible formative use of the information, but not describing a clear process for this information to inform the summative levels. This example is typical of the dataset in its description of the tests supporting teacher confidence in levelling; the tests provide a number for school data systems, whilst teacher assessment information provides more of a description. If teachers perceive a key part of summative assessment as the creation of a number or level, then tests could provide this more easily, whereas if teachers perceive the purpose of summative assessment as summarising attainment across the subject then a wider approach may be necessary. An exploration of teacher perceptions of the purposes of assessment is a further line of enquiry for the case study chapters.

4.2.3 The use of APP tracking grids

Tracking grids, which were strongly represented in the data, are a list of detailed assessment criteria or objectives which can be highlighted or ticked when the teacher judges that the objective has been achieved (or pupil if using for self-assessment), which could be done on an individual basis, with a tracking sheet for each child, or as a group. This information could be used for formative purposes, for example, to decide on the focus of the next lesson, and for summative purposes, to summarise attainment to that point. This could form part of whole school tracking systems, or there could be a separate system for tracking pupil progress of different cohorts and groups through the school for accountability purposes.

One form of tracking grid mentioned by 36% of schools was Assessing Pupil Progress (APP), introduced by the Department for Children, Schools and Families (DCSF 2010), was however no longer a government-recommended approach by the time the data were submitted to PSQM. A range of associated benefits of using the APP approach were mentioned by several subject leaders, for example:

Extract 4.5

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.

PSQM C2 reflection, March 2013 - **R4.14.4**

This extract notes the way that the tracking grid provides an 'overview' of learning, rather than a 'snapshot' which would be provided by a test. This signifies two different ways of considering summative assessment: as a summary or as a 'snapshot', which provides a further line of enquiry for the case study chapters. Extract 4.5 also mentions teacher confidence, this time provided by APP, rather than the tests noted in Extract 4.4. It could be that teachers felt supported by the explicit criteria, or it could be that the whole school initiative was supportive, with all teachers involved and following the same processes. This extract does not explain what these processes are, but states that the approach is 'consistent' across the school, which suggests some kind of agreement regarding how to use the APP tracking documents.

The following extract discusses a range of uses for the tracking grids:

Extract 4.6

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.

PSQM C2 reflection, March 2013 - **R4.23.4**

This extract lists a wide range of outcomes for use of APP: whole school tracking, profile of science, to inform planning and for differentiation. This suggests that the APP tracking grids were used for both formative and summative purposes, although the processes for each are not explained.

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. Whilst 36% of schools were using APP tracking

grids, 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 was exclusively inquiry skills-focused. Since at this time teachers were required to report attainment levels for both scientific concepts and inquiry skills it appears that there was a tendency to use separate systems for these components: typically testing for conceptual understanding and APP for inquiry skills, for example:

Extract 4.7

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.

PSQM C2 reflection, March 2013 - **R4.9.5**

This extract exemplifies the way many schools were running parallel systems, which could raise questions about manageability for primary teachers who are also responsible for classroom assessment in all other subjects. In addition, there remains the question regarding whether in reality it is actually possible to separate science into component parts, with in particular inquiry skills so ‘strongly content dependent’ (Millar 2010).

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 inquiry skills, is no longer government policy and is not compatible with the changes to the national curriculum in 2014 (DfE 2013a) – some submissions were still considering its introduction, as in the following example:

Extract 4.8

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.

PSQM C2 reflection, March 2013 - **R4.1.2**

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 trialling it. 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 removal of APP, it appears some schools found it a useful tool, and others were planning to trial it, despite their own worries, perhaps because of the lack of an alternative.

4.2.4 Summative assessment summary

In summary, for summative assessment:

- Nearly all of the schools described at least one method of assessment which could be categorised as summative, but there was often little explanation of the processes involved.
- Tests and tracking grids were the most frequent methods described, with APP particularly popular at this time.
- Over a third of the schools (37%) used a combination of methods, but it was unclear how the different information was aggregated.
- Some schools described completely separate systems for conceptual understanding and inquiry skills, raising questions regarding how the assessments separated science into constituent parts, and then how these were combined to provide an overall assessment.

4.3 Formative assessment

4.3.1 Categorising as formative assessment

In order to classify a strategy as formative assessment, a decision was required for whether the methods needed to explicitly lead to action, a key principle of formative assessment or Assessment for Learning (AfL). 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 adaptation of learning experiences. However, the subject leader reflections were brief descriptions, focused more on the listing of strategies rather than full explanation of the processes, as was found during consideration of summative assessment above. Thus, in order to map and compare methods across the PSQM sample an inclusive approach was used whereby possible formative assessment methods were coded as 'elicitation strategies', since an elicitation activity could provide assessment data without necessarily leading to actions which would support learning. 'Elicitation' largely refers to eliciting pupils' conceptual understanding (Ollerenshaw and Ritchie 1997), but for this data analysis it was used more broadly to include any potentially formative strategy mentioned by subject leaders, for conceptual or procedural understanding.

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. Nevertheless, in the listing of strategies it was unclear as to whether the strategy was being used to collect evidence, or to make a judgement, or both. Some elicitation strategies could be more focused on collecting evidence (e.g. pupil recording), whilst others could be classified as primarily judgemental (e.g. teacher marking), with others being arguably both collecting evidence and making judgements (e.g. teacher questioning or observation). For example, in recording an observation (e.g. by note-taking on post-it notes 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 any of these techniques in a science subject leader's summary is insufficient to separate it into evidence collection or judgements, all strategies have been included in the elicitation data for completeness.

4.3.2 Strategies for elicitation

A wide range of elicitation strategies were described by the 91 schools, from paper-based tasks to pupils raising their own questions, with many subject leaders listing a number of strategies, for example:

Extract 4.9

A range of assessment methods are used across the school... For example, individual and differentiated questions are used when marking for the children to respond to giving the teacher an insight into what the child has understood and how to extend their learning. Sometimes concept mapping is used at the beginning of a topic and revisited at the end when additional knowledge is added with a different colour pen. On other occasions, games, films, short play scenes and songs are produced with the known facts.

PSQM C2 reflection, March 2013 - **R4.10.2**

Sub-codes: questioning, marking, concept map, games, drama, other (songs)

Extract 4.9 provides a typical example of the subject leader listing a range of elicitation strategies for ‘planned’ formative assessment (Cowie and Bell 1999), with some just named (e.g. ‘games, films, short play scenes and songs’) and others explored in more depth (e.g. marking). It is not possible to know how embedded or prevalent the practices are from such descriptions, for whilst this extract notes that methods are ‘used across the school’, it could be that different year groups use different strategies, or that these are one-off events. For the purpose of gaining a sense of the range of strategies used across the PSQM sample, each different strategy listed by each school was tallied, but it must be noted that this does not represent the prevalence of the strategy within each school. Figure 4.3 presents a summary of the elicitation strategies present in the PSQM Round 4 data.

Figure 4.3 Elicitation strategies described in PSQM Round 4 reflections (March 2013, N=91)

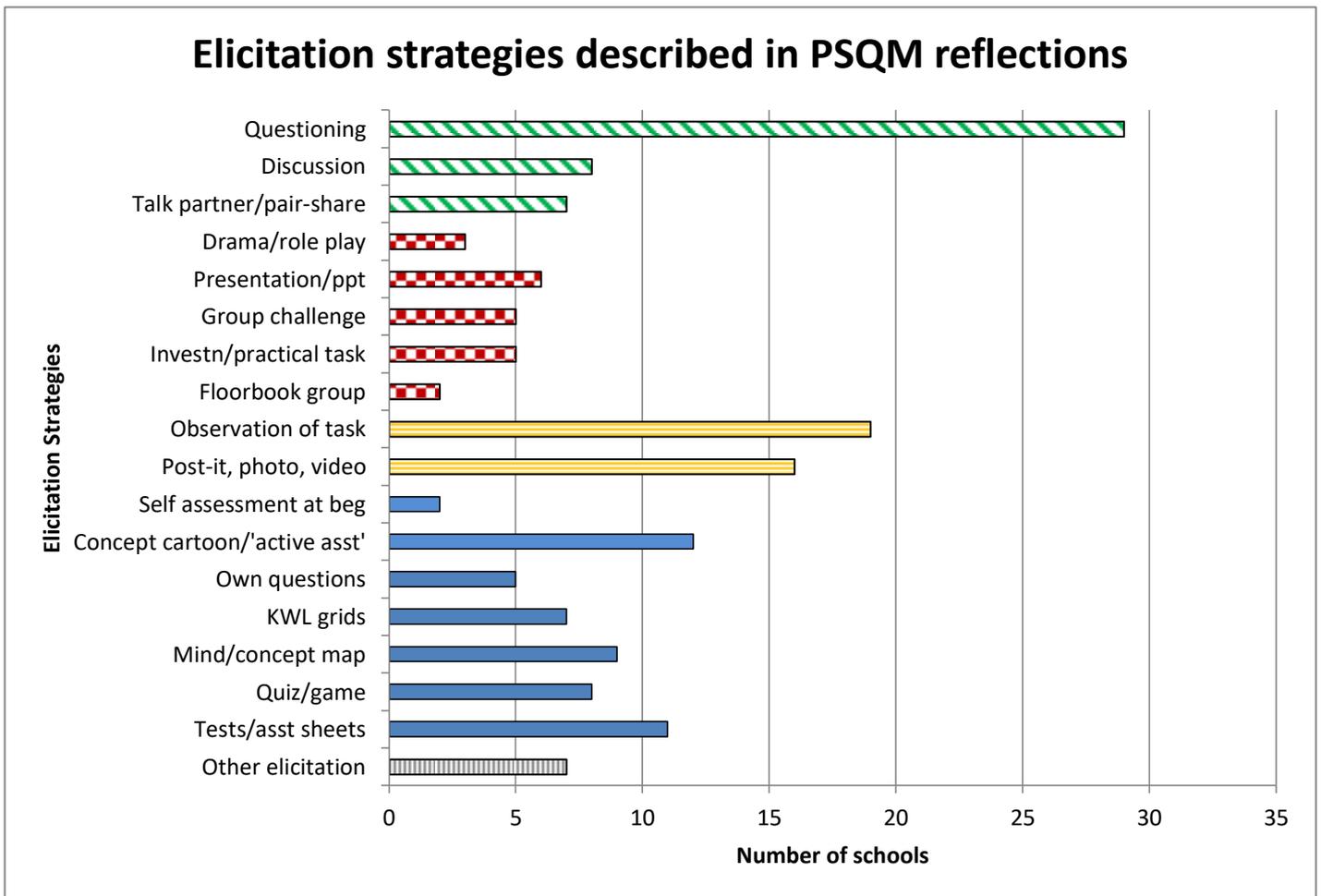


Figure 4.3 groups together similar approaches to elicitation in science such as: teacher-led talk (in green diagonal stripes), collaborative activities (in red checks), observation (in yellow stripes) and paper/task-based (in blue blocks). Some strategies were difficult to categorise from the subject 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, presentations were mentioned by five schools, but it was not clear whether the children were working alone or in a group. Thus the grouping above is to support discussion, rather than a strict categorisation.

Some researchers questioned whether schools were misinterpreting AfL to mean frequent testing (e.g. Black 2012 and Swaffield 2011), but schools in this sample, who were developing primary science, 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 support the children's learning by, for example, adapting teaching or identifying next steps.

Some of the strategies presented in Figure 4.3 merit further explanation, for example, Extract 4.10 describes two of the paper/task based strategies and how they are used:

Extract 4.10

As in other subjects, many teachers use the KWL technique at the beginning and end of science topics. It encourages the children to note down what they already know (K); what they want to know (W) and then at the end what they have learned (L). Initially, this is a valuable formative assessment tool for teachers while at the same time engaging pupils' curiosity and interest in the new topic. Ultimately, it provides teachers with information about the children's learning. In some classrooms there is a board on which the children can place questions about the topic. Again this ignites interest and gives children ownership of the topic. 'Rising Stars' is also used to access prior knowledge. In Year Two, a common misconception, highlighted in the initial assessment, was that seeds surround the root of a plant. In order to address this, the Year Two teacher bought sunflowers for children to observe the location of the seeds.

PSQM C2 reflection, March 2013 - **R4.15.1**

Sub-codes: KWL grid, own questions, tests

This extract describes a strategy in more detail than many submissions, which is why it was selected as an example here: to provide a definition of 'KWL grids'. It also explains how the elicitation strategies both provide information for the teacher and are said to engage the pupils. The formative use of assessment can be seen in the way the teacher addressed a misconception by providing an experience of seeds in sunflowers. This highlights the way tests ('Rising stars') can be used formatively, in addition to their summative role explored in the previous Section; it is the use of the assessment information which labels it as formative or summative, not the strategy itself (Harlen 2007).

Extract 4.11 provides a more typical presentation of elicitation strategies, largely as a list:

Extract 4.11

All units start with assessment of prior knowledge e.g through concept cartoons or mind mapping. Children have the opportunity to research new vocabulary and develop questions. Additional teacher assessment resources have been acquired e.g Active Assessment to complement existing ones e.g Testbase.

PSQM C2 reflection, March 2013 - **R4.10.4**

Sub-codes: Mindmap, own questions, concept cartoons/'Active Assessment', tests

Extract 4.11 includes 'Active Assessment' in its list which refers to a publication by Millgate House (Naylor et al. 2005) containing instructions for a range of elicitation strategies, including 'concept cartoons' (Naylor and Keogh 2000), which is why they are grouped together in Figure 4.3. Eight schools mentioned the use of concept cartoons, where children consider the cartoon characters' answers to a problem. Concept cartoons are designed to support discussion, so could have been grouped with the 'teacher-led talk' strategies, rather than the 'task-based' strategies, but it was not always clear whether they were used to stimulate a class discussion or for individual responses. Talk did feature strongly as an elicitation strategy, for example, seven schools mentioned the use of pupil talk partners to discuss ideas in pairs. The use of 'questioning' by twenty-nine schools is likely to have been a 'talk' strategy which could have involved individuals, groups or the whole class. 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 (William and Black 1996).

The elicitation strategies could vary in terms of how open or closed the tasks were. For example, a mind map where the child records what they know about forces could be classified as an open task whilst a true/false quiz would be deemed closed. Returning to 'questioning', this could be in the form of fast-paced closed questioning or open-ended consideration of 'big' questions such as 'what would life be like without friction?' In fact, it is likely that 'questioning' would include both convergent and divergent teacher questions (Torrance and Prior 1998). A line of enquiry for the case study chapters will be to consider the use of more open or closed strategies, to explore how divergent and convergent assessment could serve different purposes in primary science.

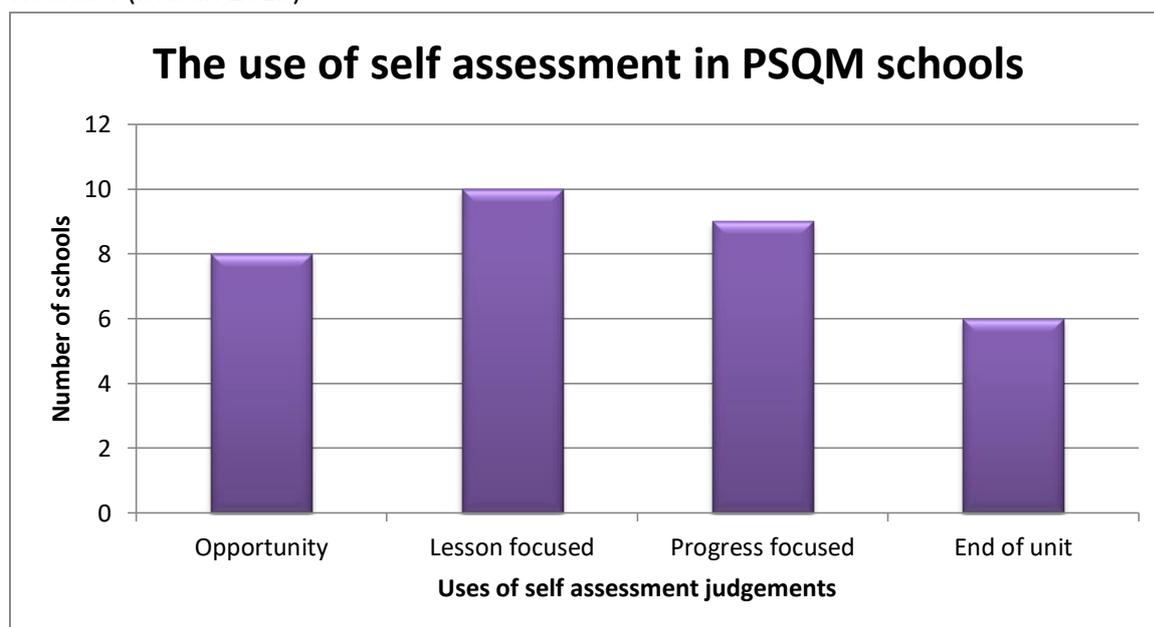
Twenty-eight schools identified feedback from teachers to pupils by marking or annotating work, although it is likely that this is an underestimation since marking is such a day-to-day routine for teachers that respondents may not have seen it as a separate assessment strategy. Exactly how 'marking' was described merits further analysis. 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 could be reduced if work was being annotated to provide evidence for accountability. Of the twenty-five schools specifically mentioning 'marking', nine emphasised teacher judgement, for example, highlighting the learning objective to show that it has been achieved. The other sixteen went on to describe how they used marking to support pupil learning 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 formative assessment is taking place, provided that children are given time to respond to the marking comments (Harrison and Howard 2009).

A further ten 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 formative assessment strategies to thirty-one - 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). Since teacher marking is such a time consuming task, it will be important to consider its use and purpose further in later chapters.

4.3.3 Pupil involvement in assessment

There is evidence in the PSQM dataset that some schools involve pupils to monitor their own learning in science. 36% of schools mentioned self-assessment and 8% peer assessment. Figure 4.4 provides a summary of the ways schools were describing their use of self-assessment:

Figure 4.4 How self-assessment was described by the 33 schools mentioning it in PSQM Round 4 (March 2013)



The first category in Figure 4.4 includes the eight schools who reported only that pupils were 'given the opportunity' to self-assess but did not elaborate, those who were more specific fell into the other three groups. Ten of the schools reported asking pupils to assess their own performance against stated learning objectives within a lesson: 'lesson-focused' self-assessment. 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. Six schools specifically said that pupil self-assessment took place at the end of the unit, signifying a more summative function. A line of enquiry during the case study chapters will

be to consider what happens to the pupil self-assessment information and whether it is utilised in teacher assessment for formative and summative purposes.

4.3.4 Formative assessment summary

In summary, the following points have emerged from consideration of formative assessment in the PSQM dataset:

- All schools listed a range of elicitation strategies, but it was often not clear whether these strategies also fulfilled a formative purpose, i.e. to improve learning. It was also unclear whether the elicitation strategies were largely for assessment of conceptual understanding or whether procedural understanding of inquiry skills was also assessed.
- Elicitation strategies included: teacher-led talk, collaborative activities, observation and paper/task-based activities. Although strategies could be grouped in different ways if further information about their implementation was obtained.
- Frequencies of elicitation strategies used by the PSQM Round 4 schools were presented, but this did not give an indication of the prevalence or embeddedness of the strategies within each school. Further evidence is needed regarding use of strategies in practice and their impact on pupils, if they are to fulfil a formative purpose.
- Just over one third of the schools mentioned self-assessment, but it was unclear whether the information gathered was used to inform teacher assessment.

4.4 The relationship between formative and summative assessment

4.4.1 Separate descriptions of formative and summative

Subject leaders in the PSQM Round 4 data devoted a considerable proportion of their reflections against criterion C2 to describing the development and monitoring of elicitation strategies in science, suggesting that formative assessment had been a focus for development in many of the schools. Summative assessment was usually described separately and in less detail, perhaps because of the uncertainty surrounding changes to statutory assessment (DfE 2013b), for example:

Extract 4.12

A concern of staff was that we were lacking any formal, summative method of assessment therefore I introduced a system of levelled tracking grids for SC1 to be completed termly: results are transferred onto our tracking system and reviewed by myself and SLT to address any trends. I delivered staff training on the implementation of these new assessment procedures and from follow-up conversations with teachers there is increased confidence in the judgements being made. Future development of formal levelling is unknown due to the changes in the programme of study; however non-levelled formative assessments are proving effective ~ accurate for tweaking unit planning/differentiation and not producing too much additional work for teachers. These will still be in place and remain useful even when levels are no longer used.

PSQM C2 reflection, March 2013 - **R4.5.1**

Summative assessment appears to be an area of ‘concern’ for subject leaders, both because they did not feel their processes were working and because of the upcoming removal of statutory level descriptors. Uncertain summative processes would be a concern for teachers in a climate of external scrutiny and accountability, with its associated need for reliability. Whilst formative assessment had a clear role, for example in this case, for: ‘tweaking unit planning/differentiation’, there was less certainty about what summative assessment would look like without levels.

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.

Ongoing teacher assessment using tracking grids, for some, was useful for providing formative information, but many felt that they needed a separate system for summative levelling, for example:

Extract 4.13

Initially embarked upon using APP grids. Whilst this was a useful tool for teachers to analyse coverage, learning and gaps it did not provide useful tool for tracking progress with sub-levels. Consequently, have decided to combine APP with more summative approach and are introducing a test scheme for years 2-6 to be included in Summer Assessment week. This will provide a breakdown of sub-levels and provide a starting point for tracking pupil progress. It also raises the importance of achievement in science in relation to literacy and maths. School can begin to build a clearer picture of standards in science across the school in terms of data. Teachers feel more confident to sub-level children and back up their own judgement.

PSQM C2 reflection, March 2013 - **R4.19.6**

In Extract 4.13, there appears to be a concern that the assessment system should produce 'sub-levels' as a way of tracking progress, so that the school has a 'clearer picture of standards in science across the school'. 'Sub-levels' were non-statutory and created by schools and local authorities to enable fine grained tracking of progress within the year. Their creation and use was one of the reasons for removing the system of levelling because it was not possible to reliably assess such small steps of linear progress (Commission on Assessment without Levels 2015). Nevertheless, the perceived need for levels and sublevels may help to explain why subject leaders in the PSQM dataset explained formative and summative assessment separately, with the former being focused on descriptive objectives and the latter needing to lead to the creation of a numerical level.

4.4.2 Links between formative and summative assessment

Some schools made links between formative and summative assessment, but on the whole, there was a lack of clarity regarding any relationship between formative and summative assessment. For example, the subject leader in Extract 4.14 described a cycle of formative and summative assessment in terms of planning:

Extract 4.14

Teachers across the school have become more and more confident in using a range of formative assessment strategies in which to assess children on a topic by topic basis. They are encouraged to use the skills of investigation as a starting point for their planning and therefore their assessment. They have been exposed to INSET to support their assessment and have been introduced to a range of resources to support them in their judgements. The impact of this has meant that teachers have become more confident in their judgements and have been able to level children successfully. This has created a loop in their assessment and teaching, enabling them to further tailor plans for groups of learners and their specific needs.

PSQM C2 reflection, March 2013 - **R4.10.3**

In Extract 4.14, the subject leader notes the introduction of a range of resources to support judgements, leading to increased confidence in summative levelling. An explicit link is made between the summative levelling and formative use, with identification of next steps. The subject leader also describes *'formative assessments'* during topics and planning for: *'the skills of an investigation'*, suggesting recognition of both conceptual and procedural understanding. Further information would be needed to explore the processes linking formative and summative assessment, for example, it is not clear whether evidence from the *'range of formative strategies'* is utilised to *'level children successfully'*. To develop a full understanding of such processes, they will need to be explored in action with practitioners during the case studies.

A different school described a range of classroom assessment strategies and resources which were used to complete a tracking grid with summative assessments:

Extract 4.15

There are a range of strategies in place for making accurate and up to date judgments on pupil progress. As a result teachers know how to take pupil learning forward. Teachers use post-its to record pupil dialogue and comments during science lessons and teaching assistants also write down relevant discussion by pupils to diagnose their understanding. Other resources used in school include Flipshare where children are filmed in action, photographic evidence, also Testbase and written work in big books. Group approach assessment is also used successfully in response to the AST visit, in KS2. The impact of this is that when assessment tracking grids are used these assessment strategies feed into this to make confident, accurate summative judgments.

PSQM C2 reflection, March 2013 - **R4.18.4**

A range of assessment evidence is listed in Extract 4.15: *'post-its to record pupil dialogue and comments', 'Flipshare where children are filmed in action, photographic evidence, also Testbase and written work in big books'*. In addition, a *'group approach'* is mentioned but not explained. The broad range of evidence suggests a sampling of a wide range of curriculum objectives, enhancing validity, but the process for *'feeding'* this information into the tracking grid needs further exploration. Nevertheless, the assessment evidence appears to be used formatively, to *'diagnose'* understanding, and summatively, to make *'accurate summative judgements'*. This appears to be more teacher-led than the process described in Extract 4.16 which used computer software called *'Classroom Monitor'*:

Extract 4.16

As a school we also use Classroom Monitor to gather our formative assessment. This has the National Curriculum learning outcomes statements for each level and each strand of science. The teacher uses judgements from their formative assessment to say if the child has met, nearly met or has not met the objective and then the programme uses this data to generate a level. (This can be over-ridden, and often is, as it usually gives too high a level).

PSQM C2 reflection, March 2013 - **R4.18.5**

The subject leader in Extract 4.16 asserts that Classroom Monitor is used *'to gather our formative assessment'*, which appears to consist of *'met, nearly met, or has not met'* judgements. However, it is unclear whether the assessment judgements described are more akin to repeated summative judgements (Black and Harrison 2010), since they appear

to serve little formative purpose. Extract 4.16 also describes the computer programme as the 'generator' of the level, which interestingly: *'usually gives too high a level'*, suggesting a lack of confidence in the accuracy of the system.

4.4.3 Separate systems for inquiry skills and conceptual understanding

The separation of scientific inquiry skills and conceptual understanding is a strong feature of the data reviewed above which is supported by 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. As noted in the literature review, although there is agreement in the literature that both conceptual and procedural knowledge should be assessed (Howe et al. 2009), the majority of recent assessment research is concerned with developing science concepts rather than inquiry skills (Hodgson and Pyle 2010, Black and Harrison 2004) and when inquiry skills have been addressed they are considered separately from concepts (e.g. Russell and Harlen 1990). 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 conceptual understanding and inquiry 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 inquiry 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.

To have separate systems for formative and summative assessment, and for the assessment of conceptual understanding and inquiry 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.

4.4.4 Relationship between formative and summative summary

In summary, the following points have emerged from consideration of the relationship between formative and summative assessment in the PSQM dataset:

- Formative and summative assessments were largely described separately, with some concerns indicated regarding changes to statutory summative assessment.
- Some schools described links between formative and summative assessment, but the processes for these were not clear.
- Many schools used separate systems for inquiry skills and conceptual understanding.

4.5 Summary and conclusions

This chapter described a wide range of formative and summative approaches reported in the PSQM Round 4 dataset. Key findings were that:

- Whilst the use of tests or tracking grids for summative assessment was widespread, few schools relied on one method alone and some described separate systems for conceptual understanding and inquiry skills. It was unclear how the different information was aggregated.
- Schools listed a range of elicitation strategies which included: teacher-led talk, collaborative activities, observation and paper/task-based activities, but it was often not clear whether these strategies also fulfilled a formative purpose, i.e. to improve learning.
- Formative and summative assessments were largely described separately, and if links were made between the two, the processes for this were not clear.

In this sample of schools there was a wide range of practice; there is *'no single approach to teacher assessment'* (Harlen 2012, 137). Whilst some schools in the sample reported using APP tracking sheets or testing, a large number used more than one method for summative assessment and this was usually described separately from formative assessment strategies. English government guidelines at the time suggested that each school should choose its own assessment structures (DfE 2013b), so such variety of practice may not be surprising. Harrison and Howard (2009) suggest that *'it is consistency of principle not uniformity of practice that works'*. Thus, variety may not be a problem, as long as methods are based on a

secure understanding of assessment purposes, which may include identifying formative or summative uses.

Of course, it is also important to remember that this sample is not representative of all English primary schools, since the sample 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 had less developed assessment practices at this time. In addition, the descriptions are unlikely to represent a full picture of practice within the sample schools: many of the schools say 'for example' which could indicate that they are not listing all of the assessment methods used within the school. The listing of strategies, rather than fuller explanation of the processes, may have been due to the C2 PSQM criterion: 'Teachers are using a range of assessment approaches'. There is also a limit to the amount of detail which can be included in a short description.

Nevertheless, the data presented in this chapter has provided a mapping of assessment approaches in March 2013, providing a useful overview to address RQ1 regarding the kinds of assessment reportedly taking place in this sample. However, the data presented in this chapter were not able to provide enough information regarding conceptualisation and enactment of the relationship between formative and summative assessment (RQ2) since the reflections listed strategies rather than provide detailed explanations of processes. In order to develop understanding, and subsequently guidance for practice, regarding the relationship between formative and summative assessment, more in-depth exploration is required in Chapters 5 and 6 through case study of the processes involved in teacher assessment in primary science.

Chapter 5

Case study A: The relationship between formative and summative assessment

5.1 Introduction

5.1.1 Chapter overview

This chapter provides an in-depth analysis of the relationship between formative and summative assessment in School A, in answer to research question 2 (RQ2). The Primary Science Quality Mark (PSQM) database discussed in Chapter 4 provided an overview of science assessment in sample schools, detailing examples of formative strategies and summative methods. Arising from this analysis were questions regarding the processes involved in teacher assessment judgements, which were not clear from the brief summaries of practice in the PSQM database. A further line of enquiry is the relationship between formative and summative assessment since they were described separately in many of the PSQM submissions. Close examination of this particular case provides the opportunity to consider practice in action, how summative judgements were constructed and how the relationship between formative and summative assessment was conceptualised.

As discussed in Chapter 3, Case Study A is an ‘instrumental’ (Stake 2006), ‘theory-seeking’ case study (Bassegy 1999); it seeks to understand how teachers could use formative assessment information to support their summative judgements of primary science. In order to explain and exemplify the processes involved, a case was selected where there was an espoused link between formative and summative assessment. The case study aimed to develop understanding of the processes involved in ‘formative to summative’ assessment, considering the way the relationship between formative and summative was conceptualised and enacted in this particular case.

The chapter begins with a brief introduction to School A and an overview of the case record which contained data from June 2013 to June 2015 (see Appendix 5A for details of the 67 items in the case record). The TAPS pyramid layers were used as an analytical framework to structure the analysis (as discussed in Section 3.7), with a view to developing understanding of the 'formative to summative' element of the framework. The discussion begins at the top of the pyramid (whole school) and moves down through the layers (reporting, monitoring, teacher and pupil layers), focusing on conceptualisation and enactment in flow of assessment information in order to track the origins of the summative assessments and how they were related to formative assessment.

The chapter will describe how School A used information gained from formative assessment in the classroom, often together with focused teacher questioning or an end of term task, to make a 'best fit' judgement of children's conceptual understanding in science. For inquiry skills they utilised detailed progressive structures called Science Stars and Skills Wheels to make success criteria explicit, enabling formative assessments to inform summative judgements. It will be suggested that such progressive structures, although not necessarily this one, support a shared understanding across the school and could be a key component in supporting teacher assessment in primary science.

5.1.2 School A context

School A is a one form entry village school in the South West of England. At the end of the case study period there were just over 200 children on roll aged 4 -11, nearly all from white British backgrounds and the number of children eligible for pupil premium (free school meals) was below the national average. It is a Church of England school which was judged by the national inspectorate to be 'good' at its last inspection (Ofsted 2012a, reference withheld to preserve anonymity).

Progress in science was described as good by Ofsted and in 2015, 100% of the Y6 children were teacher assessed at level 4 (expected level of attainment), with 37% at level 5 (above expected). The science subject leader (SL) had been teaching in primary schools for over 30 years and was Deputy Head at the school. Before the case study period she had won an

award for teaching and had produced assessment exemplification materials. During the period of the case study she became actively involved in a number of PSTT projects, began to lead cluster meetings in her local area and led her school to achieve a Gold Primary Science Quality Mark. The awards and leadership described here signify that the school took an exceptional interest in science; it was selected for case study because it had developed its own system of science assessment. The school espoused a 'formative to summative' approach to assessment, which provided a 'theory-seeking' case (Bassey 1999), to explore the way such a 'formative to summative' approach could be enacted.

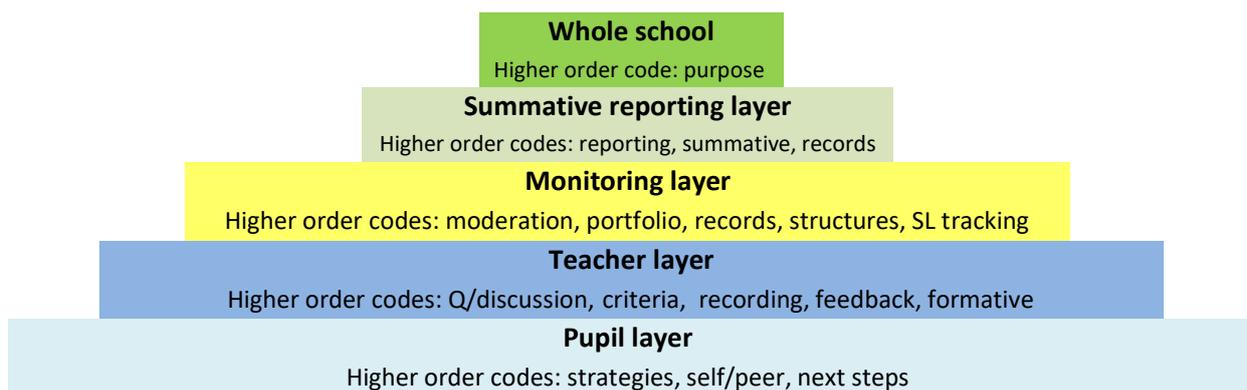
5.1.3 School A data and analysis

The data for School A were collected between June 2013 and June 2015 from 6 TAPS cluster days (discussions and written tasks), 5 school visits (non-participant lesson observations, interviews and collection of school documentation) and one PSQM application. See Appendix 5A for full details of the 67 items in the case record and a secure link to a Google Drive folder containing the source material. Each piece of data has been numbered for reference and this is included at the end of each extract (e.g. A10 for science policy). Much of the data collected (initial interview, written tasks, some school documentation and the PSQM application) was from the perspective of the SL since she acted as a representative for the school for matters concerning science. She also led the development of science across the school so her practice and beliefs were likely to influence practice across the school. In order to triangulate the reported practice of the SL, classroom observations and documentation from other teachers and classes across the school were included in the data.

All of the data collected from School A during the two year period was comprehensively analysed, but not every item is represented in the extracts below since some were focused on issues beyond the scope of this study, for example, the later school visits and TAPS cluster days were concerned with the development of other resources. Particular attention was paid to the research questions in order that the analysis should remain focused on the relationship between formative and summative assessment, whilst placing this within the rich context of the school. As discussed in Section 3.7, data analysis was supported by ATLAS.ti software which was used to organise and code all of the data collected, using

theory-led codes arising from the research questions and the TAPS pyramid, together with emergent codes arising from the data (see Appendix 5B). The ‘higher order’ codes were organised by using the TAPS pyramid layers as an analytical framework to structure the data (see Appendix 5C). The TAPS pyramid layers, rather than the more detailed boxes, are used as the framework because it is the flow of information from ‘formative to summative’, up through the layers, which is the concern for this study. The TAPS pyramid analytical framework provided a structure for systematic and comprehensive analysis of the case to consider the relationship between formative and summative assessment at each layer, from the way whole school processes were conceptualised, to enactment seen on school visits. It could be argued that separation into layers created unnatural breaks in the processes, just as coding chunks the data into parts, removing its context. In an attempt to minimise disruption of the processes, explicit links between the layers were made by cross-referencing between sections. Figure 5.1 details the pyramid layer framework and the ‘higher order’ codes for each layer, which are depicted in **bold** in the analysis below.

Figure 5.1 TAPS pyramid analytical framework: pyramid layers and ‘higher order’ codes



5.2.1 Whole school processes

The flow of information to the top of the TAPS pyramid, resulting in a ‘valid and reliable summary’ of attainment (Davies et al. 2014), is reliant upon purposeful whole school processes. The analysis begins at the top of the pyramid, the end of the process, in order to track back to the origins of the summative assessments and how they were related to formative assessment, since enactment of assessment processes is a key line of enquiry.

The top of the pyramid also contains a whole school focus and understanding of the purpose of assessment, thus extracts coded with the term '**purpose**' of assessment provided insights into the way science assessment was conceptualised and enacted at School A.

The science policy (Extract 5.1) stated that the way children are assessed: '*enables them to make progress*' and lists a range of practices which involve discussion, classwork, observation and tests:

EXTRACT 5.1

How we assess our children in a way that enables them to make progress

- *We share the learning intention with the pupils.*
- *We assess our children by talking to them and asking questions, by looking at work and by observing the children carrying out practical tasks.*
- *We use our assessments to plan for further development.*
- *Learner's work is discussed with the child.*
- *Some work is targeted for assessment purposes.*
- *At the end of Key Stage children are assessed using teacher assessment.*
- *At the end of Key Stage 2 children are assessed using NC tests.*

Science policy, collected November 2013, last updated October 2010 – **A10**

Although collected in 2013, the policy had not been updated since 2010, explaining the mention of National Curriculum tests (SATs), which many schools chose to continue to use for a short time after they were abolished in 2009. It is not clear from this policy what information was used for summative teacher assessment and how: "*some work is targeted for assessment purposes.*" A formative use was suggested when describing how assessments are used: "*to plan for future development*" and the first four points focus on children's learning, but there is no explanation of what or how this information was used for the end of Key Stage '*teacher assessment*'. Thus the policy provides limited information about assessment processes at School A, with formative and summative purposes not clearly distinguished. It also cannot be assumed that the espoused practices listed in the policy are those that occur in action, however, it is useful to look at this historical document

to provide a context for future developments and note the status which had already been given to classroom discussion, an area which will be discussed in Section 5.5.1.

When discussing the purpose of assessment with the SL (Extract 5.2), it is the formative purpose which received the most attention:

EXTRACT 5.2

The purpose of assessment in science

The purpose of assessment is to develop learning, to identify where children are, and to plan next steps. Assessment should involve children (AfL) and include some success criteria. It should also involve listening and questioning.

How the ethos of the school affects approaches to assessment

There is agreement that assessment should not be rigid or an exercise in filling in boxes. APP is not manageable. Paper testing is limiting and does not necessarily give an accurate measure of attainment.

A high emphasis is put on speaking and listening and group work which is evident throughout the school.

SL interview field notes, November 2013 – A9

Formative aims and strategies are described when asked about purpose, with no summative role mentioned; whilst summative strategies such as *'paper testing'* are listed with more negative connotations, with concerns about validity raised by the SL: *'limiting', 'not accurate'*. There appears to be both a separation of purposes and a representation akin to Harlen's (2013) 'good' and 'bad' faces of assessment. *'Rigid'* or box-filling assessments have been rejected, perhaps signifying a rejection of criteria compliance, where surface-level ticking of detailed objectives takes prominence, rather than in-depth learning (Mansell et al 2009). The primary focus on learning chimes with the literature on AfL, for example, Gardner et al. (2010: 2) similarly assert that: *"assessment of any kind should ultimately improve learning"*. Consequential validity of formative assessment requires the information gathered to be used to support learning and this appears to be the espoused focus for assessment at School A.

In order to make conceptualisations about formative and summative assessment explicit, the SL was asked to write a definition for each on the first TAPS cluster day (Extract 5.3):

EXTRACT 5.3

Formative assessment: *It is the assessment for learning which goes on all the time. The stages that you are at and the steps you can take to improve next time. The teacher is key in this - they need to ask the right questions at the right time to move the child's learning forward. Therefore the teacher needs to have a clear understanding of the steps to take along the learning journey.*

Summative assessment: *This is the end of year overall assessment based on all the assessments cumulatively or a test.*

Defining formative and summative written task, SL, TAPS cluster day 1 October 13 – A6

The SL again placed more emphasis on formative assessment, for which a more detailed explanation was provided. Timing was a key part of the difference between the SL's conceptualisations of assessment, with summative happening at the 'end of year', whilst formative 'goes on all the time'. The allocation of summative assessment to the 'end of the year' is a common way to describe Assessment of Learning (Mawby and Dunne 2012), in contrast to Taras's (2005) view of all assessment beginning with a summative judgement. This could also indicate that the SL was describing separate processes, with summative assessment seen as a separate or unusual activity, whilst formative assessment was more part of everyday practice, a process rather than an event (Swaffield 2011).

A key role of the teacher, in supporting children with their next steps, is described in formative assessment, but there is less sense of the role of the teacher in the definition of summative assessment. Summative assessment is described in two contrasting ways: a cumulative summation or a test. These are different conceptualisations of summative assessment. A summary of attainment utilises a range of information, supporting validity (Mansell et al. 2009) but the process for choice and collation of evidence to inform the summary is not explained. A test, as a 'snapshot' in time, could arguably provide less valid, but more reliable data (Halliday 2010), tipping the balance in the 'trade off' between validity and reliability in the opposite direction. This distinction between summative assessment as summary or snapshot will be explored further below in a discussion of 'best fit' (Section 5.3.1).

Any link between formative and summative is unclear, but this may have been a result of the way the task was set, asking for separate descriptions, which accentuated the split. The task also asked for definitions of the terms, which the SL gave in general terms, so it cannot be assumed that this extract describes practice in School A. The way formative assessment is described as important and valued is both a repeated and ongoing feature of the dataset, but its relationship with summative assessment is not clearly expressed. Analysis thus far has not revealed clear processes for formative assessment leading to or informing summative assessment.

5.2.2 Summary of practice at whole school processes layer

Key features of assessment arising from this layer:

- Conceptualisations of assessment included a value dimension, with summative assessment described more negatively.
- The timing of an assessment was central to conceptualisations of the difference between formative and summative assessment, with the former described for example as '*all the time*' and the latter as, for example, '*end of year*'.
- Descriptions of summative assessment as an 'overall' summary or a test could provide insight into two distinct conceptualisations of summative assessment which will need further exploration in the sections below.

5.3 Summative reporting layer

5.3.1 Summative as 'best fit'

In the 2010 school policy at the beginning of Section 5.2 (Extract 5.1), little information was given about summative assessment, other than it happened at the end of the Key Stage and included testing at KS2. By 2013 the subject leader was keen to stress that **summative** assessment arose from formative data, and that teacher assessment was valued over end of Key Stage testing (Extract 5.4):

EXTRACT 5.4

At the ends of science units a summative assessment is carried out.

There are no tests used. The summative judgement arises from formative assessment. The school used old SATs papers for a while at end of Key Stage but found these were unsatisfactory, and APP was found to be unmanageable. Therefore a whole school decision was made that summative would be informed by formative leading to a best fit model.

Summative data and planning is handed over to the subject leader and interviews with children are used to moderate progression in learning.

Questioning is the primary tool for assessing understanding at a level. This is then divided into sub levels and a best fit is derived by teacher judgement.

At the end of a topic and the end of the year, a best fit is given across all the levels.

Sc1 and knowledge form an overall level in line with National procedures.

SL interview field notes, November 2013 – **A9**

In Extract 5.4, the SL asserts that the school used a ‘formative to summative’ model of assessment, which is the reason that this case was chosen for detailed study. There is a suggestion that questioning was a key tool for the teacher to assess understanding at a ‘level’ or ‘sub level’ which appears to suggest that teachers moved from discussion with pupils to the allocation of a number based on levelling criteria. It would be wrong to assume that questioning was the only assessment tool, but it is highlighted here as a ‘primary’ one and will be discussed further in Section 5.5.1.

A recurring theme appears in this interview extract: ‘*a best fit model*’, ‘*a best fit is derived by teacher judgement*’ and ‘*a best fit is given across all levels*’. The ‘best fit’ assessment described here is where the teacher uses assessment information to find the closest match between outcomes and a National Curriculum level descriptor, as advised by the DfEE (1999) which was statutory until 2014:

In deciding on a pupil's level of attainment at the end of a key stage, teachers should judge which description best fits the pupil's performance. When doing so, each description should be considered alongside descriptions for adjacent levels. (DfEE 1999: 17)

Such an assessment could enhance validity by reducing the construct under-representation inherent in testing (Gardner et al. 2010) and enhance reliability since teacher assessment can utilise more evidence than is available through external assessment instruments (Mansell et al. 2009). However, one of the reasons for removal of the system of levels was because of the way the ‘best fit’ model produced an overall judgement which could mask gaps in understanding. The National Curriculum moved to age-related expectations whereby the lists of curricular objectives became the new criterion or ‘attainment targets’:

“Attainment targets: By the end of each key stage, pupils are expected to know, apply and understand the matters, skills and processes specified in the relevant programme of study”. (DfE 2013a: 4)

This represented a statutory shift from ‘best fit’ to full coverage, where the new expectation was that all objectives would be met; this was termed by some as a ‘mastery approach’ (Boaler 2015) or more recently a ‘secure fit’ (DfE 2017). Such a change in the way summative assessment is perceived occurred within the case study period for School A, and has significant repercussions for the relationship between formative and summative assessment, an important line of enquiry for the rest of the study.

Closer consideration of the processes of the ‘best fit’ model raises the question of how much information is actually used to inform the judgement, since there is little detail provided in the SL interview other than the mention of teacher questioning. Another source of data provides further information (Extract 5.5):

EXTRACT 5.5

The Y6 teacher was asked how she made a summative judgement:

It’s best fit, look at child’s work over term, teacher judgement about where work fits and give sublevel. Sometimes do end of term something which can be part of information, but does not ‘give’ you a level, it informs. There is no set model. Match to science levels.

Discussion with Y6 teacher, field notes January 2014 – **A12**

School A replaced judgements solely based on an end of Key Stage test, with a 'best fit' teacher assessment, drawing on a range of information which may include a '*child's work*' in normal lessons or an end of term task or question. In Extract 5.5 the Y6 teacher emphasises that the '*end of term something*' does not '*give you a level*'. This is perhaps highlighting the difference between end of Key Stage assessment procedures for different subjects, for example, the pupils sit a reading test and the score would be converted, by a pre-defined formula, into a level: the test would 'give' the level. In contrast, for a teacher assessment in science, there is no calculation or pre-defined formula, '*no set model*', to provide a 'best fit' judgement. This could be described as a strength of teacher assessment, avoiding internal reliability issues inherent in tests or tasks (Johnson 2012), and strengthening validity by using a wider range of information than a termly snapshot (Mansell et al. 2009). However, the lack of transparent processes for collating a term's work into a summative judgement, both opens teacher assessment up to criticisms of bias, especially if the judgements form part of the school's accountability measures (Green and Oates 2009), together with making it very difficult to explain the processes to others in the community. It also requires a large amount of knowledge of the subject on the part of the teacher, the teacher being entirely responsible for judging whether the pupil's answers are consistent with the teacher's 'model' or expectation of how the pupil can demonstrate understanding (Black and Wiliam 1998). Connelly et al. (2012) note that teacher judgements do more than match evidence to criteria, they draw on multiple sources of knowledge, of pupils and previous experience. Without guidance and exemplification, an inexperienced teacher may struggle to make a 'best fit' teacher assessment because they lack a clear expectation of what it would look like for pupils to demonstrate understanding in a topic, and there is a lack of transparent processes for combining such assessments into a 'best fit' judgement.

The 'best fit' model of teacher assessment was very much a model wedded to National Curriculum levels (DfEE 1999); the SL and Y6 teacher both describe above how the pupil outcomes were matched to the closest level or sublevel. As noted above, this criterion scale for summative judgements was replaced when a new curriculum was made statutory in September 2014 (DfE 2013a), with an expectation that summative judgements would no longer be 'best fit' since it was expected that all of the criteria would need to be met. This raises a number of questions about manageability of tracking for each child in each

objective, or for holistic judgements of proficiency, which will be considered further in Chapter 7. For School A, the implications of removing levels would impact on their system of ‘best fit’ because a change in the criterion scale from broad statements (DfEE 1999) to an ‘all or nothing’ list of objectives (DfE 2013a) would appear to be incompatible with a ‘best fit’ system. During the period of the case study, there appeared little structural change in the school’s systems, as seen in the Science Stars system for inquiry skills discussed in the next section, whose numbering was based on the National Curriculum levels (DfEE 1999).

5.3.2 A separate system for inquiry skills

Separate **summative records** for science concepts and inquiry skills are described (Extract 5.6 and 5.7):

EXTRACT 5.6

Staff make summative assessment for each child in Knowledge and understanding and a separate assessment for science investigation at the end of each science topic and these are then combined to give an overall Science assessment at the end of the year.

TAPS Application Summer 2013 – **A2**

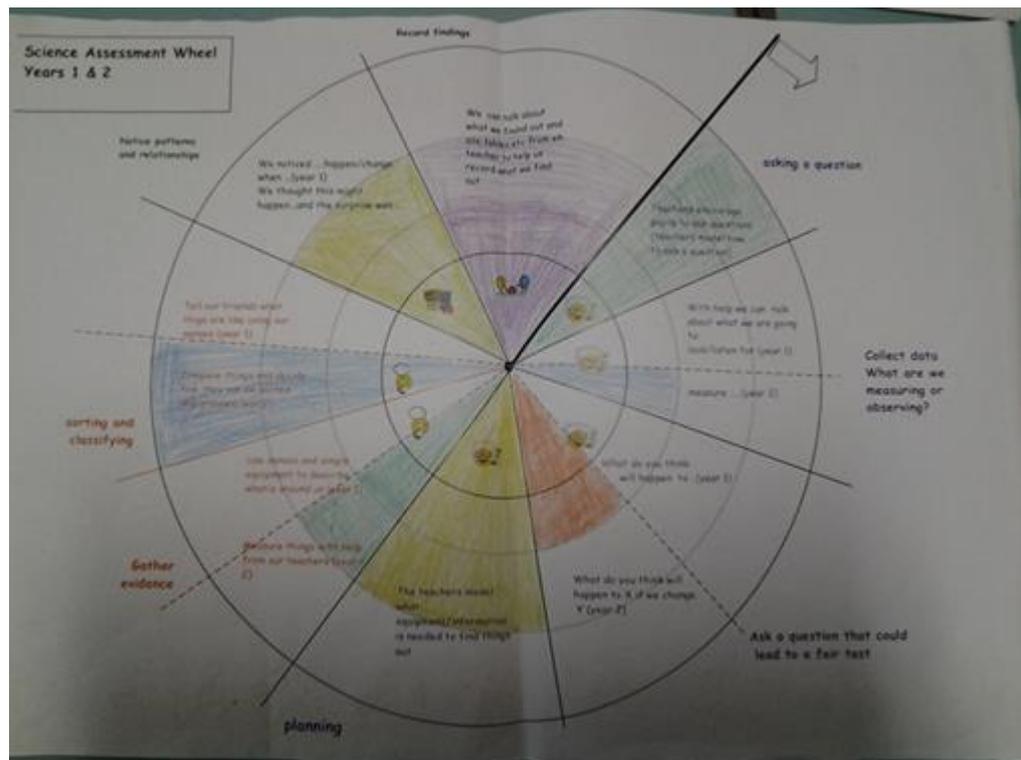
EXTRACT 5.7

A record of the skills taught is kept in the form of a Science Skills Wheel to ensure a balanced coverage of skills. Information, regarding how well the class as a whole has progressed, is passed on to inform the next teacher of the skills which need to be mastered. The Wheels, displayed and referred to during lessons have worked particularly well as the children can see how well they are progressing: they enjoy assessing how well they feel they have achieved at a skill, in agreement with the teacher.

PSQM SL reflections for C2 Assessment criterion, Spring 2014 – **A31**

The Skills Wheel described in Extract 5.7 is a summary document for inquiry skills across two year groups, for example, a Skills Wheel for Year 1/2 can be seen in Figure 5.2.

Figure 5.2 Science Skills Wheel



Science Skills Wheel, Y1 July 2014 (soon to be passed onto Y2 teacher for planning) – A42

The Skills Wheel appears to serve both formative and summative functions for the class as a whole. The SL's reflection in Extract 5.7 describes how the Wheel is displayed on the wall and referred to during lessons, to make the inquiry skills objectives explicit for formative assessment purposes (observed use during lesson observations which will be discussed in Section 5.5.2). A segment of the wheel is coloured each time the teacher judges the majority of the pupils to have achieved a particular objective within a lesson, signifying a summative purpose, albeit for the class as a whole rather than for individuals. This could also serve a formative purpose, with uncoloured objectives feeding into future planning. The collective assessment tracked class development of inquiry skills and was passed onto the next teacher, providing a both a summary of collective attainment and an identification of gaps to be addressed in the following year. School A had a one-form intake, so the class moved through the school together, thus this system provided a whole class summative record, to supplement the individual assessment information.

The school's separation of inquiry skills and content could make the assessment of primary science more manageable for the teachers, narrowing the focus onto one area for assessment and tracking. However, Millar (2010) argues that inquiry skills are both 'not well defined' and strongly content dependent; the latter view is reflected in the formulation of 'Working Scientifically' in latest National Curriculum (DfE 2013a). The school's Skills Wheel perhaps attempts to define the inquiry skills for the teachers, but this does not address the concern of content dependency. There are a number of areas for discussion: whether it is possible to extract information regarding inquiry skills from each content area or investigation context, and if it is possible, whether this is useful information if pupil performance is unique to each context. The question is not whether they can be taught separately, out of context; the National Curriculum (2013) and others (e.g. Abrahams and Millar 2008) argue that inquiry skills should not be taught in isolation, the question is whether a judgement of skill proficiency can be made. Taking a particular example from the Skills Wheel in Figure 5.2, a teacher may be able to consider pupil performance in 'classifying' across a range of contexts, making a judgement about their proficiency to classify, although they may also notice that the pupil is more confident when classifying animals than materials. The content dependent nature of science inquiry skills provides weight to the argument for the use of multiple contexts: *"several tasks are needed for a reliable assessment of any individual student"* (Millar 2010: 129), pointing towards a 'summary' rather than 'snapshot' model of summative assessment. The Skills Wheel contains the expectation of use in multiple contexts, with the colouring of one section of wedge each time. However, the tool was used for whole class tracking rather than individual assessments, so it may have masked the underperformance of individuals or groups within the class. It is possible that the 'best fit' nature of summative assessment has been transferred to this tool, so that it displays a 'best fit' judgement for the class, which can then inform whole class planning.

The class Skills Wheel provided general information for the next teacher about class inquiry skills, but individual description is not contained in this document and it could be questioned whether this document is more about tracking coverage rather than attainment. The individual records passed to the SL and next teacher were at this time in the form of numerical levels (**A43**), for which one could again question whether hidden behind the

number would be areas of strength and weakness. A question is once again raised about the amount of detail and description for individual children of both inquiry skills and content: is the purpose of summative assessment to provide a summary or a detailed record of the child's performance in each area? This school appeared to be passing on a summary, in terms of levels and class Skills Wheel, with more detailed outcomes contained in pupil work books. It would appear unmanageable for class teachers to keep detailed records for each child, but it also appears important to recognise that learning behaviour from a range of inquiries should inform summative assessments of inquiry skills. However, the atomistic fragmentation of inquiry skills remains a concern (Ollerenshaw and Ritchie 1993) making it important to examine practice in lessons to determine whether the Skills Wheel was used atomistically or as a 'bridge between atomism and holism' (McMahon and Davies 2003: 37) for focused teaching within the context of a full investigation.

5.3.3 Summary of practice at summative reporting layer

Key features of assessment arising from this layer:

- Summative assessment was conceptualised as a summary, enacted as a 'best fit' judgement which aimed to draw on a range of information but which also may mask gaps in attainment.
- There is no set process for making a 'best fit' judgement and it may rely on in-depth teacher knowledge of the subject; which may make it difficult for inexperienced teachers, without further guidance and exemplification.
- Separate summative judgements were made for concepts and inquiry skills, raising questions for how science was broken down atomistically and then recombined into a holistic judgement.
- A Skills Wheel structure was used to record class performance and this is passed onto the next class teacher, which appeared a manageable way of tracking and informing whole class planning, but it was unclear how this related to the assessment of individuals.

5.4 Monitoring layer

5.4.1 School structures

In addition to the Skills Wheel discussed above, School A utilised a range of school **structures**, to support and monitor the curriculum, particularly, the County scheme of work, which listed objectives and key questions, and the school's Science Stars, which contained the levelled criteria from the Skills Wheel in 'child speak'. The Science Stars are essentially a child-friendly version of the National Curriculum level descriptors, associated with symbols and key vocabulary. The appropriate stars were displayed in classroom, for example, in a year 5 class the children were aiming to be '4 star' or '5 star scientists'. In the PSQM submission the SL described the importance of these school structures, pictorially in the portfolio (Figure 5.3) and explained in more detail in the reflection (Extract 5.8):

Figure 5.3 School structures

The figure is a collage of educational materials. On the left is a large circular 'Science Assessment Wheel' with multiple colored wedges. Below it is a text box: 'Science Assessment Wheel on display in Y1. The teacher colours in the inner wedge when the skill has been taught and outer wedges when most/all children in the class have mastered the skill.' To the right is a display board titled 'Examples of Science Stars on display' with a note '2* in Y1 and 5* in Y6'. The display board features several cards with handwritten notes and drawings, such as 'FT M R', '5* +', '5* Predict what might happen to the Dinosaur when you change a variable', and '5* Interpret Explain Analyse Line graphs and Tables'. At the bottom right is a starburst containing a student's reflection: 'The Science stars help me with my learning because if I get a mark in my book I can look up on the display to see how I can improve next time.' The number '3' is visible in the bottom right corner of the collage.

PSQM portfolio, assessment slide, Spring 2014 – A26

EXTRACT 5.8

A range of assessment practices are carried out, including annotation of photos, concept maps, SATs style questions, and Concept cartoons. However, the 'County' Scheme of Work provides the basis of our assessment: it includes a series of questions helping teachers (and TAs) discover the levels children are working at and how to help children aim higher with their learning. The questions help in lesson planning and Science Enquiry skills are assessed through "Science Stars," ranging from 1-5*(corresponding to levels). These are made explicit to the children, and displayed in the classroom. The stars are given verbally in class and through the marking in books, particularly at KS2.*

PSQM SL reflections for C2 Assessment criterion Spring 2014 – **A31**

In Extract 5.8, the SL suggests that the questions on the scheme of work supported:

"teachers (and TAs) [to] discover the levels children are working at and how to help children aim higher with their learning". This indicates use of the structures for both criterion-referenced summative judgements of levels and formative next steps, although it is unclear whether this is part of the same process. The list of strategies (*"annotation of photos, concept maps, SATs style questions, and Concept cartoons"*) are not clearly linked to outcomes or purposes of assessment in this extract. This appears to contradict the: *"no tests are used"* comment from the SL interview (Extract 5.4 in Section 5.3.1) since SATs-style questions are test questions. This could be resolved by noting the difference between a whole end of year test which provides the entirety of the summative judgement, and the use of SATs-style questions which could form part of the judgement. The SL asserts that Science Stars are *"made explicit to the children"* through displays, discussion and in marking, which was borne out on school visits and would support the 'shared understanding' characteristic of the Monitoring layer. A question is raised regarding whether it is the teacher or pupil who is the primary user of the Science Stars: the teacher uses the Science Stars to match pupil outcomes to the criteria, but it is unclear whether the pupils also use the criteria; this will be discussed further when exploring self and peer assessment in Section 5.6.

On school visits the Science Stars and Wheels structures were visible on the walls of the classrooms, and were referred to in lessons. They appeared to provide a structure for

teachers and children to consider the inquiry skills focus for the lesson and where this fitted into the inquiry cycle (Wheel) or into a progression of inquiry skills (Science Stars), both developing a shared understanding of the nature of science inquiry and how to improve. If discussed in a lesson, for example to clarify success criteria or next steps, these could fulfil a formative purpose. If the Skills Wheel or Stars are used as criteria for a 'best fit' judgement, these structures are supporting a summative assessment. There is perhaps a 'mid-way' use, where the children or teachers are using the structure to judge their attainment in the current lesson, which is perhaps where formative and summative purposes align. The same evidence can be judged formatively and summatively (William and Black 1996). However, by placing a numerical judgement on the evidence there is a risk that it is seen by pupils as the end of the process (Black and Harrison 2010), an area for discussion which will be revisited in Section 5.5.3 when considering marking.

The explicit **structure** and **criteria** provided by the scheme of work, the Science Stars and the Wheel appear to be supportive for staff (Extract 5.9 and 5.10):

EXTRACT 5.9

Teacher new to school in January, in previous school no science assessment. Found here that Scheme of Work really useful - bands of progression. Plans include level statements, show what aiming for eg L3, Qs, egs of misconceptions

Y4 teacher comments after lesson observation, School visit 3, March 2014 – **A36**

EXTRACT 5.10

As a result of CPD training and moderation staff meetings over the last 2 years staff also report feeling more confident at providing a higher level of challenge. "The Science Wheel helps me focus on areas I haven't taught and the scheme of Work clearly shows me the next steps for learning" (Y1 teacher). As a consequence, staff have become much keener to try investigations of their own.

PSQM SL reflections for C2 Assessment criterion, Spring 2014 – **A31**

The structures appeared to be used for multiple purposes: supporting the planning of next steps and identifying gaps in the teaching ('*areas I haven't taught*'), together with supporting teachers to understand '*what [they are] aiming for*'. The multiple uses of the

school structures, to support both formative and summative assessment, could support manageability in the system. Alternatively, it could lead to a confusion of uses, just as using the same assessments for multiple purposes can confuse their function (Gipps 1994). For example, the scheme of work's level statements could narrow the teaching so that attention is only paid to that which will 'tick the box', or the Science Stars could promote atomism during inquiries as the teacher narrows the focus onto the one area that is not coloured in yet. How the structures work in practice will be considered further in the Teacher layer in Section 5.5.

5.4.2 Moderation

Staff confidence in using the inquiry skills structures and in levelling of conceptual work was supported by staff meeting **moderation** sessions (Extract 5.11):

EXTRACT 5.11

Science moderation meetings are now a regular feature on the staff meeting agenda. For example, every 2 weeks during the Autumn Term, the meeting began by looking at some Science work brought along by a member of staff. The concept is identified and a level agreed. When this was introduced initially 2 years ago, the process took a whole meeting but, as staff have become more skilled, the moderation exercise now takes approximately 10 mins. The work samples have been collated into a Science assessment portfolio, available for reference in the staff room. "The moderation meetings have really helped me. I am far more confident at levelling and it is useful to have the portfolio available to refer to," Y4 teacher. The moderation meetings have also meant that Science is regularly kept on the staff meeting agenda and I can feed in updates/reminders whilst enabling me to monitor teaching and learning across the school – a very powerful tool to keep the profile of Science high in the school!

PSQM reflection B1 Spring14 – A30

The SL noted that the 10 minute moderation staff meeting slots had been developed over a 2 year period, which represented a significant time commitment for the school, showing the importance placed on science at School A. The SL describes an increase in staff confidence in making summative judgements of pupil work, which could have been supported by: the regular 'drip feed'; the 2 year commitment; the support from the school structures or

subject leader; the way staff took it in turns to share a variety of work; the discussion and process itself; or the development of a portfolio.

In Extract 5.11 the purposes of the moderation meetings are described as: raising staff confidence at levelling; keeping science on the agenda or keeping the profile high; and enabling monitoring of teaching and learning across the school. However, beyond '*a level [is] agreed*', there is little mention of reliability, which for assessment theorists is a most important role of moderation discussions (Johnson 2013). Harlen (2007) argues that whilst teacher assessment is often perceived as having low reliability, with effective moderation procedures, the reliability of teacher assessment can be as high as it needs to be in the 'trade off' between reliability and validity (William 2003). School A appeared to be using moderation staff meeting discussions to serve multiple purposes, to both check judgements and as a means of staff development (Green and Oates 2009). The SL's reflections appear in line with the Connelly et al. (2012) findings which recognised both the usefulness of peer support for developing consistent judgements and the amount of time and effort involved in the complex process. Although the detail of the judgement making processes are not described, by regular demonstration and participation in the moderation discussions, the teachers appeared to gain confidence in making summative judgements on pupil work drawn from classroom activities, which could indicate use of 'formative to summative' assessment.

Extract 5.11 above notes the compilation of a **portfolio**; this is explained further in an assessment exemplification document created by the School A (Extract 5.12):

EXTRACT 5.12

We decided to make a portfolio of levelled and annotated science work to help teachers develop their assessment for learning skills as well as for children, parents and Governors. We wanted it to show progression through the school and to encourage staff and children to see the next steps and to “aim high”.

Staff wanted support with science assessment, so a series of 10 minute science moderation slots took place within staff meetings. Staff took it in turns to share a piece of work and we agreed a level. As a result, a moderation file was set up with the examples we generated and this developed into a more visual record with photos (as Ofsted loomed!)

Collecting a portfolio of Science work is a great way of celebrating and sharing the range of investigations going on across the school. Moderating regularly in small manageable chunks helps us to maintain a high profile for science, gives teachers confidence and means we have super evidence of children’s attainment.

The portfolio will be updated and adapted to provide examples of assessment to match the new National Curriculum 2014.

PSTT Assessment exemplar 2013 – **A4**

The school were clearly advocating the creation of a portfolio of levelled work, however, the planned update for the new National Curriculum did not happen, and the SL was unable to find the portfolio during the school visits. It could be that the process of creating the file, with staff meeting discussions of different types of science work from different year groups, was the important factor for developing staff confidence, rather than the folder as a reference tool. Portfolios were described as a key strategy for supporting summative assessment by Black et al. (2011), but for School A, the portfolio was a staff development tool rather than a portfolio of evidence for a particular child, as in the study. The recording made by pupils will be discussed further in Section 5.5.3, suffice to say that in School A, portfolios were not constructed for individuals, but for examples of work which had been levelled by staff.

A question arises about how the levelling of individual pieces of work in the moderation staff meetings, for inclusion in the portfolio, relates to the ‘best fit’ judgements discussed in Section 5.3.1. For teacher assessment to enhance validity, it needs to be based on a range of information. It is not clear how the pupil work discussed in the moderation meetings fed

into summative assessments, since the judgements were on individual pieces of work. The portfolio and moderation meetings were used as staff development opportunities, with the focus more on making judgement processes explicit and developing teacher assessment literacy (Edwards 2013), rather than on making summative judgements about particular children. Harlen (2007) asserts that it is the evidence which should be used for summative judgements rather than an aggregate of the scores; it is not clear how the agreed levels on the work would be used by staff when making summative judgements.

5.4.3 Summary of practice at monitoring layer

Key features of assessment arising from this layer:

- School A utilised structures to support teaching and assessment across the school: Science Stars, Skills Wheels and a scheme of work containing level descriptors and key questions. These structures were used both formatively and summatively for criterion-referenced assessment.
- The structures supported staff planning, science coverage and assessment confidence. They appeared to help build a shared understanding of science assessment across the school.
- Regular moderation took place in staff meetings where pieces of work were levelled. This appears to have been a key method for developing a shared understanding of the science criteria and the school assessment processes. Moderation made the process of making summative judgements explicit and could provide an example of 'formative to summative' assessment.
- A portfolio of levelled work was developed as a result of the moderation staff meetings. It was suggested that it was the process of making the portfolio, rather than the product itself, which supported conceptualisation and enactment of summative assessment.

5.5 Teacher layer

5.5.1 Teacher questions

The SL described how assessment strategies, particularly teacher **questions**, were built into the planning structure used by the school (Extracts 5.13 and 5.14):

EXTRACT 5.13

Staff use the 'County' scheme of work to plan and assess. They build in assessment opportunities through careful questioning/observations/mind maps/ simple tests.

PSQM self-evaluation Summer 2013 – **A1**

EXTRACT 5.14

We use the 'County' scheme of work which included AfL questions at all levels to help staff assess the science objectives being taught to inform their next teaching steps and to help the children know how to progress to the next level in their learning.

TAPS Application Summer 2013 – **A2**

The 'County' scheme of work provided objectives and key questions for teachers to ask, but did not contain detailed lesson plans, so teachers used the scheme to inform their day-to-day planning rather than use it as a recipe to follow. Below is an example of key questions included in the scheme (Extract 5.15):

EXTRACT 5.15

Level 1: Tell me what is happening to these materials when we put a magnet next to them?

Level 2: Sort these materials into groups. Those that are magnetic and those that are not.

Level 3: Which of these metals are magnetic? What happens when you put the 2 ends of the magnets together?

Level 4: How can you measure the pulling force of a magnet? What happens when you put like poles together, and opposite poles together?

Key questions for magnetism from 'County' scheme, collected July 2014 – **A38, A39**

These key questions are largely prompts for the pupil to show and explain to the teacher what is happening when magnets and different materials are brought close together. Such

questions could be used for formative or summative purposes, especially since the questions are ordered or 'levelled'. The lesson observations described below provide further information regarding classroom discussions.

5.5.2 Discussion and use of criteria in lessons

Analysis of conceptualisations of the relationship between formative and summative assessment have so far been based on documentation and interview material, but to consider enactment in more detail it was necessary to observe classroom practice. This section will explore the use of classroom **discussion** and the use of assessment **criteria** in lessons, by close analysis of observed practice in three lessons. For these three lessons the categories from the TAPS pyramid teacher layer (Davies et al. 2014, drawn from Harlen 2013) were used as an observation schedule (later observations were more focused on the development of resources so are less relevant to this discussion). A brief overview will be given to provide the context for each lesson and summary table 5.1 supports comparison of the observation field notes, organised by categories within the TAPS pyramid teacher layer.

The subject leader's Year 5 space lesson on School Visit 2 (**A11-12**) began with a whole class carpet discussion about the Earth and sun. In the main part of the lesson the pupils worked in pairs or threes to physically model the orbit of the Earth around the sun using different sized balls. As the children moved the Earth ball they gave a commentary on what was happening, which was then peer-assessed for clarity and accuracy. On the same day, parts of a Year 6 lesson on inheritance were observed (**A11-12**) with carpet, paired and table discussions around vocabulary, family characteristics and matching dog breeds. On School Visit 3, a Year 4 lesson on branching keys (**A35-36**) was observed; this involved the children working in threes to create a branching key to sort animals by writing yes/no questions on post-its.

Table 5.1 Lesson observation field notes organised using TAPS pyramid Teacher layer criteria

Lesson	Y4 Keys lesson Creating post-it keys to categorise animals	Y5 Earth in space lesson Using balls to model orbit of Earth	Y6 Inheritance lesson Exploring inherited charact in dogs and own families
Date	March 2014	January 2014	January 2014
Data identifier	A35, A36	A11, A12	A11, A12
Teachers involve students in discussing learning goals and standards	<i>Raised hands to show if find keys tricky. Mini-plenary to look at others' work – what do you notice?</i>	<i>Importance of using science vocab 4 or 5* scientists</i>	(did not observe start of lesson)
Teachers gather evidence of their students' learning through questioning/discussion	<i>Discussed kind of Qs in branching database. Open Qs for talk partners: What hab in sch? What is it like? – asked for more detail</i>	<i>Qs emphasising expl - probed explanations and meaning/use of vocab - Withhold judgement so ch have to expl for selves</i>	<i>Probed children's meaning of inheritance vocab 'No hands up' strategy</i>
Teachers gather evidence of their students' learning through observation	<i>Gps building postit keys – spotted clearest and pointed ch in that direction</i>	<i>Observe groups modelling Earth orbiting</i>	<i>Pairs recording ideas on whiteboards while T circulates</i>
Teachers gather evidence of their students' learning through study of products	<i>Post it branching database Asst notes on plans for ch that stand out – above or below</i>	<i>Look at group's explanation and modelling. Draw and write explanation.</i>	<i>Whiteboards to note family characteristics. Written expl of what learnt and examples</i>
Teachers use assessment to advance students' learning by adapting the pace, challenge and content of activities	<i>Previous lesson found branching keys difficult so doing in mixed ability groups Pupils identify how to improve their key eg missing Y/N</i>	<i>Physically modelling Earth's orbit in a circle since virtual expt looks like oval. Did not move onto day and night since challenged enough by orbit whilst spinning.</i>	<i>Provides word to help expl eg characteristics, structure for recording: Mum, Dad, me.</i>
Teachers use assessment to advance students' learning by giving feedback	<i>Go around grps to check on clarity of Qs</i>	<i>Asking if can use better science words,</i>	<i>Say more than 'face' for characteristics</i>
Teachers use assessment to advance students' learning by providing time for students to reflect on and assess their own work	<i>Evaluating Qs Pairs walk around and look at other's work – what notice? Return to own keys and improve</i>	<i>4th child in group to listen and watch – are they explaining using sci vocab, watch groups and give feedback, decide if 4 or 5* scientists and wr in margin</i>	(did not observe end of lesson)

The use of Science Stars to make success **criteria** explicit, a key feature of formative assessment (William 2011), was observed in the subject leader's Year 5 lesson on space (Table 5.1, column 3, row 4). The groups gave advice to each other for how to improve their explanations. The teacher emphasised that to 'become 5 star scientists', they should aim to use scientific vocabulary accurately, which led to the pupils listening out for the word 'orbit' or 'axis' in the explanations. In contrast, the field notes for the other lessons did not contain reference to Science Stars. It appears that the Science Stars structure was more embedded within the classroom practice of the SL, or that she was drawing attention to them to demonstrate their use to the observer. The Science Stars were visible on the walls in all three classrooms and some numerical scoring was seen in both the Y5 and Y6 pupil books (A13-A21), which will be discussed further in Section 5.5.3 on recording and marking.

The Science Stars appeared to be a structure which could be used in class with children, when marking or when planning, but they were not apparent in every lesson observed. The SL used them formatively, to make explicit the success criteria of using scientific vocabulary, and summatively, asking the children to decide on the star rating of their explanation at the end of the lesson (Table 5.1, column 3, final row). The Science Stars structure could provide a bridge between formative and summative, using the same criteria for both, or it could be more in line with Taras' (2005) view that each assessment begins with a summative judgement. It could also be questioned whether such tightly focused expectations, for example, the focus on accurate use of particular vocabulary for '5* scientists' described above, whilst supporting reliability, could potentially reduce validity. The tight focus may lead to other aspects being ignored, or perhaps a 'tick box' culture where the teacher is waiting for a particular word like 'orbit' rather than probing understanding across the topic. The pupil's role in this will be discussed further in Section 5.6.

Classroom **discussion** was a prominent feature in all three of the observed lessons summarised in Table 5.1 (row 5). Each teacher used strategies like talk partners to increase participation and wait time (Rowe 1972). For example, in the Y6 lesson, pair talk dominated with the teacher 'listening in' to discussions to support her formative assessment, then asking probing questions to stimulate further discussion. In the Y5 lesson, it was noted that the teacher was '*withholding judgement*' during the class discussion about

the Earth and sun (Table 5.1, column 3, row 5). The teacher questioning focused on explanations and use of vocabulary, but the teacher did not say 'that's right' and move on. This supports a more dialogic (Alexander 2008) approach to discussion, moving beyond the mere 'call and response' of interactive-authoritative dialogue (Mortimer and Scott 2003). By withholding summative judgement of pupils' answers, the children were prompted to explain further and the teacher received richer formative assessment information, from a greater number of children.

The prominence of talk in all three lessons supports the school's espoused policy (A10). Of course, these are just three lessons and the presence of an observer is likely to have affected how the teachers behaved in the lessons, but it is interesting that all three chose to present a lesson so full of dialogue as 'best practice' to an observer. The ephemeral nature of such lessons could be a problem for reliability, since traditionally only evidence in permanent form receives attention for summative assessment (Black and Wiliam 1996). It would not be manageable to record every pupil utterance, yet a lack of recording could result in a biased view of pupil attainment, for example, only those confident in their answers may be willing to share their ideas, masking the less confident or those with difficulty in accessing the concepts. Some of the strategies used by the Y6 teacher could alleviate this, for example, the recording on mini-whiteboards and the use of 'no hands up' for answering questions (Table 5.1, column 4, row 5-6). Both the Y5 and Y6 teacher ended the lesson with the pupils individually recording their ideas in writing (A15-19), which provided not only a permanent source of evidence to utilise for a later summary assessment, but also an individual record, in contrast to the rest of the lesson which had been work in pairs or trios (Table 5.1, column 3, row 6). Recording will be discussed further in the next section.

5.5.3 Recording and marking

In a 'formative to summative' approach, what pupils record and how this is marked are important decisions since these become sources of evidence for making summative judgements. In the observed lessons described above, pupil **recording** consisted of: 'throwaway' post-its, mini whiteboard notes, or a brief recording at the end; raising

questions about whether such outcomes could contribute to a summative judgement made some time later. The SL's comments on pupil recording are presented in Extract 5.16:

EXTRACT 5.16

SL comments on recording:

Encouragement is given to the recording of thinking and skills. Children have a science note book to record their ideas.

TAs and teachers are encouraged to scribe/record children's utterances. Recording – higher and lower attainers are given more emphasis.

Children often orally communicate their understanding which is captured in floorbooks. All systems are geared towards children knowing where they are and whether they can communicate this to others.

SL interview field notes November 2013 – **A9**

The SL notes that pupils are encouraged to record their ideas and adults are encouraged to scribe pupil's utterances, sometimes capturing this in a 'floorbook' (a large-format, 'home-made' book – **A56**). Such recording is: '*geared towards children knowing where they are*', suggesting a role for pupil self-assessment, which will be discussed further in Section 5.6.

There was no explicit guidance on whether particular types or times of recording should be given more prominence for assessment. The Y6 teacher's explanation of 'best fit' in Section 5.3.1 described how a teacher would: '*look at child's work over [the] term*' (Extract 5.5, **A12**) whilst flicking through a child's work book. The record made by the child over time supports validity since it is based on a range of different activities and contexts. However, validity may be compromised by a lack of recording in the more practical or discussion based lessons, for it is less likely that the teacher will use this ephemeral evidence in their summary judgement. The Y4 teacher commented after the lesson that she makes: '*Assessment notes on plans for children that stand out – above or below*' (**A36**), just as Extract 5.16 notes: '*higher and lower attainers are given more emphasis*'. Such teacher annotations on planning provide a manageable way of gathering further evidence to inform later summaries, although this would only be for a limited sample of the class. Noting children that '*stand out*' requires teacher knowledge of what 'expected performance' looks like for the assessment criteria, and this is where the less-experienced Y4 teacher found the

school structures useful, for example, the scheme of work whose: *'plans show what aiming for e.g. L3' (A36)*. School structures could support teacher understanding of the assessment criteria, enhancing reliability, together with enabling ongoing assessment during classroom discussions, formatively to focus questioning, and summatively to note individuals on planning.

Teacher **marking** provides a significant opportunity for formative and summative assessment (Extract 5.17):

EXTRACT 5.17

SL comments on marking:

Children are given the opportunity to respond to marking at the beginning of sessions.

Next steps are built into this and the Sc1 wheel is a visual aid.

Skills are communicated to children and a star rather than a level is used as success criteria.

There are explicit discussions with children about levels/star symbols and science skill wheels.

Marking is used to feed judgements back to children.

SL interview field notes November 2013 – **A9**

The SL describes how marking provides opportunities for both providing a teacher judgement and a 'next step' which may require a response from the pupil at the beginning of the next session. Evidence of both of these was seen in books, for example, marking to the objective with teacher questions and pupil responses (**A20-21**), together with numerical Science Stars which were seen in the margins of the Y5 and Y6 books (**A13-14, A20-21**). The inclusion of both numerical judgements and comments is reminiscent of Butler's (1988) study which found that the inclusion of scores cancelled out the positive effect of feedback via comment-only marking. Black and Harrison (2010) also argue that the score signifies that the process is complete; the judgement has already been made. They recommend that comment-only marking is used, with any scores recorded for the teacher's use only. This would appear to fit with the school's structures which support teacher understanding, although it is unclear whether the numbering system also supports a teacher 'shorthand' which is useful for tracking and summarising. Teachers may summarise the scores rather

than the evidence, which Harlen (2007) warns against; however, re-reading each piece of work from each child at the end of the year sounds an unmanageable expectation. A further line of enquiry for the next section is how much the assessment criteria, with or without numbers, should be shared with the pupils.

5.5.4 Summary of practice at teacher layer

Key features of assessment arising from this layer:

- The scheme of work contained levelled teacher questions which could support elicitation of pupil understanding for both formative and summative purposes.
- Three observed lessons demonstrated a high status given to class and group discussion, which provided a rich bank of ephemeral evidence, but may be difficult to draw upon for summative assessment without some form of recording or note-taking.
- Science Stars criteria were displayed on each class wall and utilised for formative and summative assessment in one observed lesson.
- Individual recordings were made in pupil work books or floorbooks. This evidence may provide the basis for the summative ‘best fit’ judgements.
- Marking is largely comment based, but there is some use of summative numerical Science Stars, which raises questions about how summative judgements are shared with children.

5.6 Pupil layer

5.6.1 Self and peer assessment

In the Nuffield (2012) pyramid model of ‘formative to summative’ assessment, the base layer was focused on teachers; TAPS added a pupil layer to recognise the active role pupils could take in assessment (Davies et al. 2014). This section will consider the role of pupil assessment in the ‘formative to summative’ approach. **Self** and **peer assessment** were observed in both the Y4 and Y5 lessons described in Table 5.1. For example, in the Y4 lesson the children worked in threes to create a branching identification key to sort animals by

writing yes/no questions on post-its (A35-36). The pupils had struggled with keys in the previous lesson and the teacher used this information formatively to adapt her planning in order to revisit the task. After talk partner discussions to raise questions (e.g. Does it have four legs? Does it eat meat?), the children were asked to self-assess their confidence in making a branching key; this was used formatively to create mixed-confidence groups. Peer assessment was also used formatively during a mini-plenary in the middle of the lesson, when pupils were asked to walk around to look at each other's keys and to pick out elements of a successful key before returning to improve their own key.

In the Y5 space lesson the children were encouraged to clarify the task and assess their confidence by discussing the meaning of vocabulary and which parts they found easy or difficult (A11-12). Peer assessment was a key formative strategy when groups watched each other's modelling and gave feedback for how to improve their explanations of the Earth's orbit. At the end of the lesson pupils were asked to draw and label how the Earth moves around the sun, then self-assess more summatively by deciding if their explanations were 4 or 5 'Star'.

In both lessons peer assessment was used formatively by the pupils in the monitoring of their learning: the pupils were asked to peer assess each other's work, provide feedback and had time to act on that feedback to improve their modelling or keys. The use of the feedback within the lesson is significant because it makes the assessment fully formative; the pupils are activated as resources for each other (Wiliam 2011). The timely feedback and improvement also makes the formative assessment manageable within the lesson. However, it could be questioned whether such highly valuable formative assessment can contribute to summary judgements, being both largely ephemeral and more pupil than teacher focused. The 'formative to summative' model, where data is reduced in transition between layers, does not need to draw upon all formative information; which formative information is most fruitful will be a line of enquiry for later chapters.

Both lessons featured explicit **criteria**, which supported reliability of pupil and teacher judgements. In the Y4 lesson, the class constructed the success criteria for what constituted an effective branching key within the lesson. Whilst in the Y5 lesson, the children made

suggestions for the features of a successful explanation of the Earth's orbit, but this was guided by the use of Science Stars, which the SL suggests are designed to support self-assessment (Extract 5.18):

EXTRACT 5.18

The investigative objectives are made clear to the children through a "science stars" display in each classroom which outlines what they need to learn to be good scientists and these help them to reflect on their achievements and to self – assess.

PSQM SL reflections for A2 Principles criterion Spring 2014 – **A27**

The school's system of Science Stars provides explicit criteria for the progression and a common language for staff and pupils to structure feedback, supporting reliability. However, as discussed above, the allocation of a numerical value to a piece of work could distract the learner from the formative feedback and signify the end of the process, as well as possibly leading to labelling of the learner rather than the work.

Self and peer assessment, particularly the latter, have been used formatively in the lessons observed, but there is little evidence that this information was used to inform summative assessment. The moderation staff meetings were focused on development of teacher assessment (Section 5.4.2), and the key questions provided in the scheme of work to support assessment judgements were provided for teacher use (Section 5.2.1). It appears that at School A, self and peer assessment serve a primarily formative function.

5.6.2 Summary of practice at pupil layer

Key features of assessment arising from this layer:

- Explicit science focused objectives and criteria were shared or developed in lessons, supporting reliability of teacher and pupil judgements.
- Self and peer assessment in lessons were used primarily for formative purposes, raising the question of whether all formative information needs to be utilised in a 'formative to summative' approach.

5.7 Conclusion

5.7.1 Key features of assessment practice at School A

Using the TAPS pyramid layers as an analytical tool (layers in bold below), the following key features of science assessment in School A have emerged:

- **Whole school processes:** espoused priority given to formative assessment, conceptualisations of assessment included a value dimension.
- **Summative reporting layer:** Summative assessment was conceptualised as a summary, enacted as a 'best fit' judgement which drew on a range of information; separate systems were in place for inquiry skills and knowledge.
- **Monitoring layer:** school structures like Science Stars and the Skills Wheels provided explicit criteria for staff and pupils, for criterion-referenced formative and summative assessment; regular moderation discussions have supported the development of a shared understanding of the science criteria and the school assessment processes.
- **Teacher layer:** a high status was given to class and group discussion but written information may be more likely to be selected to inform summary judgements; teacher questioning was supported by the scheme of work; pupil recording and annotated planning may have provided the basis for summative 'best fit' judgements.
- **Pupil layer:** self and peer assessment used for formative purposes only.

School A self-evaluated their assessment processes using the TAPS pyramid tool and selected green for the top of the pyramid, denoting that in science they felt that they provided a valid and reliable summary of student achievement (**A33**). School A summative assessment could be described as reliable because there are explicit criteria in the scheme of work, Science Stars and Wheels, which provides a structure for shared understanding and moderation discussions. The moderation sessions in staff meetings were highlighted as a way to support reliability in assessment, however, once the PSQM was completed and with the advent of a new National Curriculum for all subjects more staff meeting time was devoted to English and maths, leaving little for science moderation. Questions around the manageability of moderation procedures were already raised above when discussing how the profile of levelled work had not been updated.

The 'best fit' summative assessment could be described as valid because it was based on a range of evidence from across the subject, including the outcomes from classroom activities where the primary purpose may be formative. However, it has been argued in this chapter that the processes for 'formative to summative' judgements were not explicit and that there were at times different conceptualisations of summative assessment. Discussion and peer assessment appeared to be primarily formative strategies, whilst the 'writing up' in books appeared to be information which could be used for summative judgements. It has been suggested that the use of numerical Science Stars could 'close down' the formative dialogue, for example when marking in the pupil books, and there was little evidence that self or peer assessment was utilised in summative assessment. In addition, the separate structures for concepts and inquiry skills could lead to questions of validity with regards to atomistic teaching, an issue we will return to in Chapter 7. It could also be questioned whether the Science Stars could provide a valid assessment of the latest curriculum since they were based on the old National Curriculum levels (DfEE 1999). In response, the school were updating their Scheme of Work to match the new National Curriculum 'age-related expectations' (statutory from September 2014, DfE 2013a), however, the embedded 'age-independent' levelness of the Science Stars remained. Nevertheless, the school structures made the criteria explicit, providing a common language and a shared understanding for staff and children. These structures could be used to support both formative and summative assessment, providing success criteria within the lesson and criteria for summative judgements. It is perhaps these common criteria which provide a bridge between formative and summative assessment, providing opportunities for the same classroom activities to be used to inform both formative next steps and summaries of learning.

5.7.2 Tentative generalisations on the relationship between formative and summative assessment

Tentative or 'fuzzy' generalisations (Basse 1999) provide a way of describing factors which were seen in the analysis of School A which may be useful for other schools. The 'fuzziness' acknowledges the uniqueness of each context, whilst also noting that there can be features

which have relevance for other contexts. Tentative generalisations which can be drawn from the case study of School A suggest:

- Teacher conceptualisations of formative and summative assessment may include value and timing dimensions.
- Summative judgements can be conceptualised as a snapshot end of topic event or a summary of progress. A 'best fit' judgement is an example of the latter, which can be based on formative assessment information, summarised for summative purposes.
- Explicit criteria provide opportunities for criterion-referenced formative and summative assessment. Teachers can be supported to use explicit criteria with school structures like a scheme of work or inquiry skills progression chart (in this case a Wheel and Stars).
- In order to use a range of information for summative judgements, teachers may draw upon pupil recording and notes from pupil discussions (e.g. annotated planning).
- Numerical systems may be useful for teacher tracking, but may signify the end of the formative process for pupils.
- Moderation discussions and school structures can support teacher assessment literacy and a shared understanding of the subject.
- High quality dialogue and self/peer assessment are valuable for formative purposes, and may not necessarily be utilised in summative assessments.

School A provided an example of a school in which classroom assessments were used for formative and summative purposes, supported by explicit criteria, school structures and moderation discussions. However, questions have been raised regarding the way structures and processes within the school have been stable through a time of statutory change, for example, maintaining a 'best fit' numerical system rather than moving to a 'secure fit' version of summative assessment. It also appeared that even in this exceptional case, where there is confidence in science assessment, the assessment processes and relationship between formative and summative assessment were not explicit.

With a pre-existing system of 'formative to summative' assessment, School A was selected as a 'theory-seeking' case (Bassegy 1999), to explore the processes, conceptualisation and enactment of such a system. Chapter 6 will provide a case study of School B which was selected as a 'theory-testing' case (Bassegy 1999) in order to explore changes over time during the implementation of a system of 'formative to summative' assessment.

Chapter 6

Case study B: Changes to the relationship between formative and summative assessment

6.1 Introduction

6.1.1 Chapter overview

This chapter presents a three year case study of School B in order to explore changes over time in the conceptualisation and enactment of the relationship between formative and summative assessment, in answer to research question 3 (RQ3). The case study of School A presented in Chapter 5 explored the relationship between formative and summative assessment processes in a school which claimed to be using a ‘formative to summative’ model, but it provided little insight into how such a system developed. The three year case study of School B, covering the first phase of the TAPS project, provides an opportunity to consider the processes and implications for changes to assessment practices during the development of a ‘formative to summative’ model.

Chapter 5 presented tentative generalisations (Bassegy 1999) which will be explored further in this chapter: conceptualisations of assessment in terms of value and timing; summative assessment as an attainment summary or ‘snapshot’; the use of school structures and moderation to support criterion-referenced assessment; the types of information utilised for summative summaries; and the role of pupils in assessment. This case study will further explore these findings, together with identifying other areas pertinent to this case, in order to develop recommendations for practice regarding the use of ‘formative to summative’ assessment.

As in Chapter 5, in order to provide a comprehensive analysis of changes in conceptualisation and enactment of the relationship between formative and summative assessment in science, the TAPS pyramid layers are used as an analytical framework (Davies et al. 2014). In order to compare changes across time, the 3 year period is split into 3 Design-Based Research (DBR) Phases: Exploration (1), Development (2) and Implementation (3). Thus the sections below consider each layer of the analytical framework through each of the DBR Phases.

The chapter will describe how, during the Exploration Phase (1), School B's concerns were in terms of reliability: for evidence, consistency, marking and standardisation. During the Development Phase (2), the teachers took part in moderation discussions and trialled strategies, considering their manageability and whether they were able to provide evidence. During the Implementation Phase (3), the data indicate developments in validity of assessments, with more confident teacher assessment, more open tasks, more active pupils and the use of assessment information for formative and summative purposes. Summative assessment came to be conceptualised as less of a 'bolt on' and more of an attainment summary, informed by a range of information collected formatively. During the case study period, conceptualisations of, and the relationship between, formative and summative assessment is developed, although balancing the demands of each was at times problematic and the case indicates additional guidance may be needed for implementation of a 'formative to summative' model.

6.1.2 School B context

School B is a small but growing village school in the South West of England. In 2013 there were 5 classes, with a mixed Year 3/4 class and mixed Year 5/6 class; by 2015 this had risen to 7 classes, one per year from Reception to Year 6. In 2015-16 there were 183 children on roll aged 4 -11. Nearly all children had English as their first language and the number eligible for pupil premium (free school meals) was below the national average. The similarities between School A and B support comparison between the assessment processes of each. It is a Church of England School which was judged by the national inspectorate to be 'good' at

its last inspection (Ofsted, 2012b). Although there was no mention of science in the Ofsted report (2012b), teacher marking was singled out as an area for improvement:

What does the school need to do to improve further?

Raise achievement in writing and mathematics from good to outstanding by strengthening teachers' marking so pupils see links to their individual targets and know how they can improve their work.

(Ofsted 2012b, reference withheld to preserve anonymity)

The school's Key Stage 2 teacher assessment results indicate that pupil attainment was high, relative to the national average. In 2015, 100% of the Y6 children at School B were teacher assessed at level 4 (expected level of attainment, national average 89%) and 75% were graded at level 5 (above expected level of attainment, national average 40%).

The science subject leader (SL) had taught in primary schools for over 20 years. At the beginning of the case study period (March 2013), she submitted a Silver Primary Science Quality Mark application which explained how science had been developed across the school. During the period of the case study she began to lead cluster meetings in her local area, won a teaching award and led her school to achieve a Gold Primary Science Quality Mark (March 2015) which signified impact beyond the school.

The school's sustained involvement with both PSQM and the TAPS project indicates a commitment to the development of primary science, which makes this three year case study possible, but also suggests it is not a 'typical' school, in which science is a '*poor relation*' to English and maths (Wilshaw 2016). Similar to the case study of School A, the case study of School B is also 'instrumental' (Stake 2006), whereby interest in the case is driven by an outside concern, to identify the relationship between formative and summative assessment over time in order to develop recommendations for practice. Bassey's (1999) categorisation of a 'theory-testing' case study can also be applied, since School B could also provide a 'test' of how the development of the TAPS pyramid influenced assessment of primary science within the school over time.

6.1.3 School B data and analysis

School B was selected from the TAPS project group because it provides the most complete case record for changes over time to be explored, being one of the few schools which did not have a change of Head teacher or science subject leader (SL) during the project. The data for School B were collected between March 2013 and June 2016 from 8 TAPS cluster days (discussions, written tasks and SL presentations), 6 school visits (non-participant lesson observations, interviews and collection of school documentation) and two PSQM applications (see Appendix 6A for full details of the 86 items in the case record). The boundary for the case is science assessment within the school, with data from across the school utilised, for example, lesson observations, pupil work samples and observation of a whole staff moderation meeting. Nevertheless, the large majority of the data comes from the perspective of the SL who was representing the school at TAPS cluster day interviews and writing the PSQM submissions. It is acknowledged that there may have been a tendency for the SL to present a more positive picture, for example, the PSQM reflections were written with the aim of securing an award for the school. However, the SL's viewpoint about what constitutes a 'positive picture' for assessment in primary science provides additional information regarding conceptualisations of assessment. The viewpoint and reported practice of the SL were triangulated by classroom observations, planning and work samples from across the school, together with comparison of sources over time, for example, between the two PSQM applications or between SL interviews or presentations.

For this case study, it was change over time which was of particular interest and the DBR phases provided a structure for this, helping to compare early analysis with later analysis. For example, the DBR phases allowed checks between the frequency of codes for a particular time to avoid over-emphasis on the 'loudest or brightest' data (Cohen et al. 2011). Deciding where one phase should stop and another should begin was a challenge because the school development processes did not stop and start, they were continuous. Thus the split into DBR phases designated breaks which could have affected the way the data was analysed and interpreted. However, the phases did relate to key events in the cycles of development, for example, the sharing the first version of the TAPS pyramid with participants in February 2014 marked a change from Phase 1 exploration to Phase 2 development.

The case record was organised into the three DBR phases as detailed in Table 6.1. The data identifier (utilised in Appendix 6A and summarised in Table 6.1) contains the phase (e.g. Ph1, Ph2 or Ph3) to support navigation, and the phases will be referred to below by their number to avoid repetition of the date range each time.

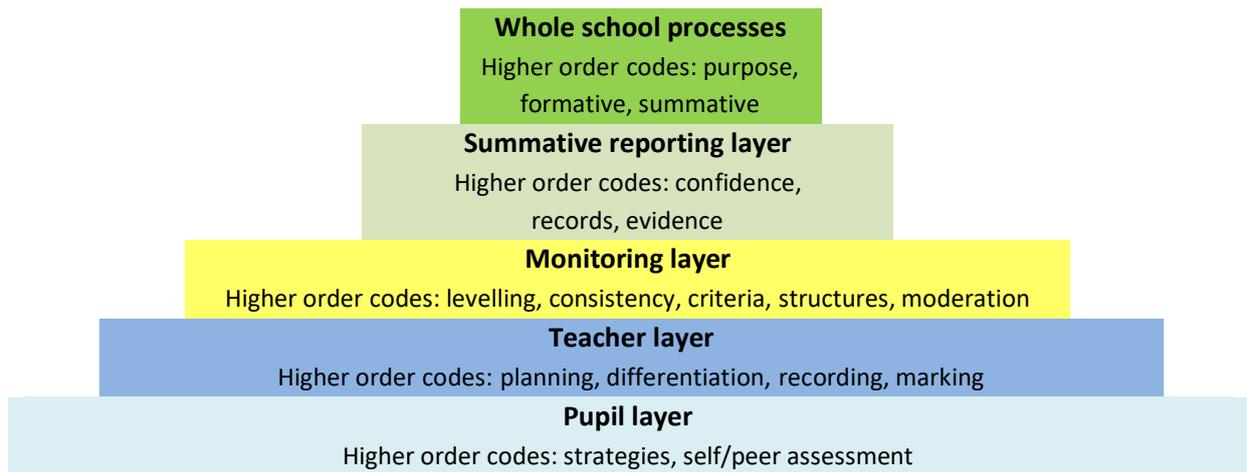
Table 6.1 Design-Based Research Phases

	Dates	Data identifier
DBR Phase 1 Exploration	March13 – Nov13	B1-Ph1 to B21-Ph1
DBR Phase 2 Development	Feb14 – Jan15	B22-Ph2 to B53-Ph2
DBR Phase 3 Implementation	March15 – June16	B54-Ph3 to B86-Ph3

The entire case record was interrogated using ATLAS.ti in batches according to the DBR Phases to allow for comparison over time, for example, comparing the frequencies of codes at each phase which could represent shifts in focus for the SL or school (Appendix 6B). For this case study the aim was to look at practice over time and an inductive approach was taken in order to strengthen the ‘voice’ of the school. Codes were added to ATLAS.ti as they emerged in the data, rather than beginning with a predesignated list as used for case study A. Nevertheless, there is no suggestion that the analysis was a ‘grounded theory’ approach, where codes, themes and theories emerge from the data without a pre-conceived framework, since the TAPS pyramid (Davies et al. 2014) supported both coding and thematic analysis.

After coding of the full case record, the TAPS pyramid layers were used as an analytical framework, in a similar way to case study A, to structure the data and identify ‘higher order codes’ (further details in Appendix 6C). Figure 6.1 details the pyramid layer framework and the higher order codes for each layer.

Figure 6.1 TAPS pyramid analytical framework: pyramid layers and ‘higher order’ codes



In the discussion below the ‘higher order’ **codes** for case study B are written in bold on their first occurrence in a section, to support transparency of data analysis. Use of the TAPS pyramid layers, as for case study A, supported a comprehensive mapping of school processes and consideration of the data from a number of perspectives. For example, ‘**strategies**’ is considered in both the teacher and the pupil layers, and ‘**evidence**’ in a number of layers as a recurring theme. Extracts have been selected to support the most relevant section in order to avoid both repetition of extracts and the use of short snippets which lack context; more extended quotations are presented to provide ‘thick description’ (Geertz 1973).

In order to support close analysis of School B’s assessment processes and any changes during the three years, the case study will consider each layer of the analytical framework in turn, over time e.g. the Monitoring layer during DBR1, DBR2 and DBR3. This enables a focus on one strand at a time, for example, within the Pupil layer, from trialling strategies in DBR1, to developing new strategies in DBR2, to developing the role of the pupil in DBR3. The case study will begin at the top of the pyramid, in a similar way to case A, in order to track the origins of summative assessment and how much they are informed by formative assessment, together with identifying conceptions of the purpose of assessment, which will impact on practice at all other layers.

6.2 Whole school processes (W)

6.2.1 DBR Phase 1W - Formative or summative purpose

What is understood by formative and summative assessment will drive the whole school processes, which aim to result in a 'valid and reliable summary' of attainment (Davies et al. 2014). This section will explore the way the relationship between formative and summative assessment was conceptualised near the start of the case study period when the TAPS project began.

Three key members of staff attended the first TAPS cluster day in October 2013. Each were asked to individually record their understanding of **formative** and **summative** assessment; their responses are presented in Table 6.2.

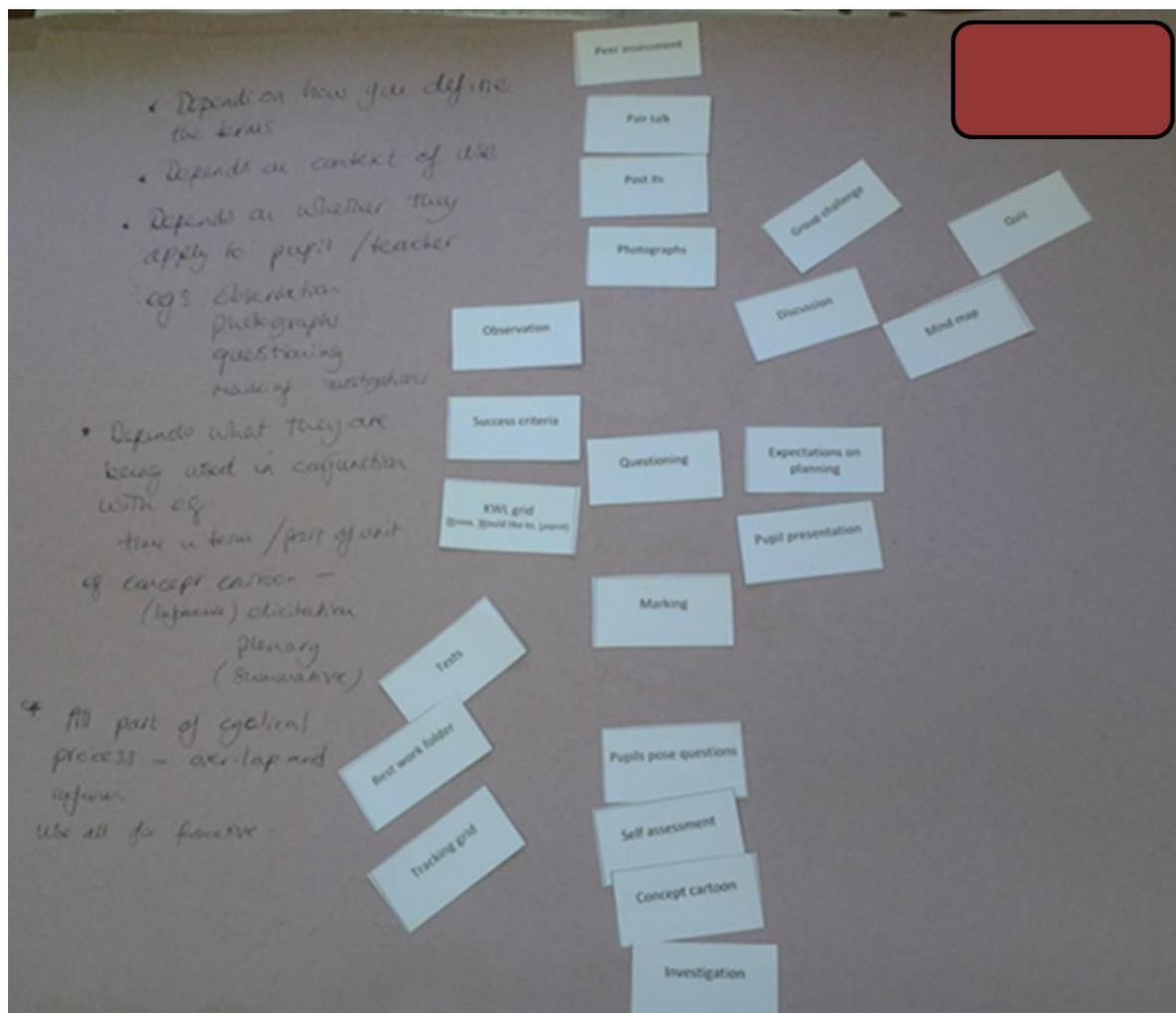
Table 6.2 Formative and summative written task, TAPS cluster day 1 October 2013 - B5-Ph1

Head teacher	Science subject leader	ICT subject leader
What does 'formative' assessment mean to you?		
<i>Ongoing - day to day within the classroom. Assessment that informs your teaching/planning and next steps in learning for class, groups, individuals</i>	<i>Purpose of formative is to inform me where children are 'at' in their learning in order to ascertain what the next steps are. This takes many forms – observing whilst they are working, listening to discussions between pupils, questioning them directly, prompting etc. Examining work and responses within activities.</i>	<i>Ongoing - every day, every lesson. Informs planning, teaching, learning. Communication - colleagues (especially job share and TAs). Feedback to children: marking/verbal. Various methods.</i>
What does 'summative' assessment mean to you?		
<i>End of unit assessment - capturing an end point. What has been learned? What progress has been made? Currently capturing a summative level. Data tracking. Testing (where/when appropriate).</i>	<i>This is matching pupil performance/attainment against national/external criteria set as benchmarks to measure and qualify performance</i>	<i>Data. Tracking. Testing as appropriate/required A necessary evil. Various methods.</i>

Responses for formative assessment include repeated mention of: *ongoing, informs teachers' next steps, many forms/methods*. There appears to be a consensus amongst the three members of staff from School B that formative assessment is an ongoing, daily activity which provides information for the teacher; as noted in Chapter 5, timing is a key dimension of the assessment conceptualisation. The SL explains that formative assessment could take a range of forms and be utilised to plan a pupil's next steps, whilst summative assessment was more about criteria matching. Repetition in the responses for summative assessment include: *numerical data for tracking, includes tests where appropriate, capturing 'an endpoint' or 'pupil performance' which can be compared to 'external criteria' or a level*. The Head teacher describes summative assessment as *'capturing a level' or 'endpoint'* suggesting summative assessment is viewed as an end of term snapshot (Mansell et al 2009), rather than an attainment summary, a key difference also identified in Case Study A. The ICT subject leader describes summative assessment as *'a necessary evil'*, in line with Harlen's (2013) findings that teachers viewed formative as 'good' and summative as 'bad', providing support for a value dimension to assessment conceptualisations.

These descriptions of formative and summative assessment appear to be quite separate, but the task asked them to be described separately, potentially accentuating the differences rather than any relationship between the two. In a different activity on the same day, formative **purpose** appeared to come to the forefront. A list of strategies for assessment (e.g. *questioning, tests, observation, self-assessment*), drawn from the PSQM data discussed in Chapter 4, were presented as a card sort (Appendix 6D) and in school groups the teachers were asked to sort the cards 'with formative and summative in mind'. The task outcome for the three members of staff from School B is pictured in Figure 6.2.

Figure 6.2 Strategies sorting card activity (with majority of strategies placed in the middle suggesting they could be used formatively or summatively)



Writing next to sorting activity: *Depends on how you define the terms. Depends on context of use. Depends on whether they apply to pupil/teacher eg observation, photographs, questioning. Depends what they are being used in conjunction with eg time in term/part of unit eg concept cartoon – (informative) elicitation, plenary (summative). All part of a cyclical process – overlap and inform. Use all for formative.*

Strategies sorting card activity, TAPS Cluster Day 1, October 2013 - **B6-Ph1**

During the card sort task and its accompanying discussion School B appeared to feel as though they should be sorting the cards in a particular way, perhaps into separate lists or Venn diagrams with an overlap for both formative and summative, like some of the other schools in the room. Their written notes repeatedly use the word 'depends', noting the purpose, context, timing and meaning of the words as criteria which must be known before

the card sort can be completed. This activity has also been used at a range of TAPS dissemination events (Appendix 6E) and has often provoked a similar response with teachers first trying to sort the strategies into separate categories, and then coming to the conclusion that each strategy can be used for formative or summative purposes, that it is the purpose which defines its categorisation rather than the strategy itself.

School B's final comment in Figure 6.2, to 'use all for formative' appears to suggest a dominance of formative purpose, but little link with summative: the relationship between formative and summative assessment is unclear in the Phase 1 data. The means of data collection could be strongly influencing the outcome, with the school attempting to match their response to what they perceived would be the 'right answer', focusing on the formative 'good' side of assessment (Harlen 2013).

6.2.2 DBR Phase 2W - Formative for summative purpose

In the card-sort activity above (Figure 6.2), the teachers asserted that all of the strategies could be utilised for formative purposes. During DBR Phase 2, a wide range of strategies were seen in use on school visits (**B22-Ph2**), but a concern for **evidence** emerged, in which it appeared that the teachers were looking for ways to use the information gathered during **formative** assessment, in order to make their **summative** judgements. For example, in the moderation staff meeting, concerns were raised about evidence:

Extract 6.1

What's required to level a child?

Evidence, range from different sources, consistently and securely, including talk- do we need a broader range of strategies to collect info from talk?

Doing science differently eg drama.

Dilemma eg if missed out on unit of work, or not get all done especially if use Classroom Monitor [tracking software] – does it aggregate?

Need to look for more strategies for talk, opportunities, peer talk comparing 2 things, do we need more ways to record this?

Self or peer assessment criteria labels tried at beginning but became redundant because still had to assess work but did raise their awareness, SC1 toolkit.

Need to look for evidence gathering strategies.

Moderation staff meeting field notes, June 2014 - **B40-Ph2**

During the staff meeting the teachers were exploring what they needed to be able to *'level a child'*, to ascribe a summative grade. The suggestion of utilising a *'range of different sources'* and looking for ways to capture oral pupil talk, could enhance validity, with the summative judgement based on information which was gathered using different instruments and representing a range of constructs within the curriculum (Mansell et al 2009). The emphasis on evidencing such judgements could be related to a concern for reliability, with the teacher providing examples (**B41, B42-Ph2**) so that the judgements could be checked by others – a concern for inter-rater reliability (Black and Wiliam 2012). Such practice appears to be in line with Nuffield (2012) and TAPS pyramid (Davies et al. 2014) recommendations that information gathered for formative purposes could be summarised for summative reporting. However, the formative purpose appears to be lost in this extract, subsumed by a concern for evidence, with each assessment opportunity becoming a summative assessment (Taras 2005). There appears to be a very fine line between summarising formative assessment and repeated summative assessment, and it will be important in the final chapters to consider how to avoid the loss of formative purpose. The collection and use of evidence will be explored further in Section 6.3.

6.2.3 DBR Phase 3W - Formative and summative purpose: assessment as ongoing

The replacement of levels by new National Curriculum criteria (published September 2013, statutory from September 2014 for Y1, 3, 4 and 5, statutory for Y2 and 6 from September 2015) necessitated a change in summative practices, but the change in criteria appeared not to be a key focus for the SL. When asked about how assessment had changed at School B as a result of the TAPS project, the SL described how summative assessment has grown less formal and dependent on published resources, whilst understanding of formative assessment had developed.

Extract 6.2

In what ways has your school changed the ways you assess children's progress in science as a result of the TAPS project?

Much less formal/summative/published material – as a bolt on or end of

Greater understanding of AfL – exploring different strategies to trial (ok to abandon)

Teacher questionnaire, TAPS cluster day 6, June 2015 - **B76-Ph3**

In Extract 6.2 the SL describes the school's previous use of resources to support summative assessment as 'a bolt on', suggesting that summative assessment had been seen as a separate entity to formative assessment. Degree of formality and degree of separation from typical classroom activities appear to be two further dimensions to teacher conceptualisations of assessment. Increased trust and confidence in teacher judgement is reported, but it is not clear how this relates to formative and summative assessment. A fuller explanation was given during an interview the following year, by which point the SL was reporting the use of formative assessment information for summative purposes:

Extract 6.3

What changes to science assessment have you made across the school in the last three years?

We don't attempt to buy any summative-type add-on assessment strategies or publications, so we don't ask for the purchase of books that will do end-of-unit assessments. We don't look for outside material. We are purely making judgments based on our own assessment-for-learning practices informed by references to other places but not added on.

*The main change is that our assessment is ongoing. We don't do any summative testing at the end of unit, so at the end of the year, we are continuously gathering data, more information about the children that informs a consensus of an idea at the end in terms of offering our head teacher or our management a summative grade. Then the rest of the time, our assessment for learning is about that: Where are these children now? Where do they need to go next? We've very much embedded assessment for learning, we do very little of the summative type stuff, and we have found the best assessment activities are the best teaching * (5:15—unclear).*

Transcribed SL interview, June 2016 - **B83-Ph3**

Formative assessment, *'embedded'* AfL, is described as *'ongoing'*, and this is reportedly used to make summary judgements at the end of each unit and year, informing a *'summative grade'*. The description of *'ongoing'* assessment signifies a move away from the view of summative assessment as an end of term snapshot; assessment as a process rather than an event (Swaffield 2011). The flow of information from formative to summative is described and there are clear purposes for summative assessment in passing it on to the next teacher and for tracking. In this extract there is less description of formative purpose, assessment seeming to be about *'gathering data'* to inform summative judgements, which could indicate a criticism of the TAPS pyramid with its flow of information serving the summative purpose; this will be explored further in the next section below.

6.2.4 Summary of changes at whole school processes level

Key features of changes in assessment practice in this layer:

- In DBR Phase 1 conceptualisations of assessment mirrored findings from Chapters 4 and 5: formative and summative were described separately, with key dimensions including timing and value judgements. Degree of formality and separation from typical classroom activities were identified as further dimensions for teacher conceptualisation.
- The relationship between formative and summative assessment was not clear at the beginning of the case study period, although teacher responses were perhaps more influenced by data collection methods at this time.
- Over time there became an increasingly strong focus on gathering evidence for summative assessment, raising concerns whether *'formative to summative'* was conceptualised as repeated summative assessment rather than a summary of formative assessment. Closer examination of enacted practice is considered in the next section.
- Later conceptualisations of assessment indicated a focus to *'embed AfL'*, with the continuous gathering of data from formative assessment to inform the *'summative grade'* so that summative assessment was no longer a separate *'bolt on'*. This indicates a move towards a *'formative to summative'* model, but it is important to note that this was not until the third year of the case study.

6.3 Summative reporting layer (S)

6.3.1 DBR Phase 1S and 2S - A focus on records and evidence

This section will explore the school's description of **evidence** and **records**. A focus on records and evidence for assessments was present in DBR Phase 1, before the TAPS project began:

Extract 6.4

- *Teachers are developing a variety of methods to gather evidence of learning e.g....*
- *Formative assessment is largely based on Teacher assessments. To try to 'standardise' summative assessments, some published material is used, including past SATs papers. Our teachers are using a range of materials to inform judgements including: T R SoW, QCA material, APP, Bucks County Council 'Level criteria/Doing Science', SATs, Rising Stars – and other materials found on internet...*
- *Teachers keep records and some annotated evidence of their assessments, which they use to inform over-all judgements at the end of each year.*
- *Pupil attainment is reported to the SL and HT for tracking purposes 3 times a year. Summative levels are recorded at end of KS – with priority weighting given to Sc1. KS2 reports sub-levels. Data reported to governors.*

TAPS application June13 - **B3-Ph1**

The SL describes the development of: *“a variety of methods to gather evidence of learning”*, which could enhance validity, sampling across the curriculum, together with possibly enhancing reliability by lessening the impact of issues with individual tasks (Johnson 2012). Teachers' records and annotated evidence are used: *“to inform over-all judgements at the end of each year.”* The school also describes how multiple published materials or **structures** were used to try to *'standardise'* their summative assessments, indicating a concern for reliability. However, the focus appears to be largely on summative assessment; it is not clear from this extract whether School B saw formative assessment as purposeful itself or whether it was primarily done in order to gather information for summative purposes.

It seems that the teachers have an 'evidence gatherer' role (Gipps et al. 1995), they do not feel confident to make a summative judgement without a large amount of supporting paper evidence, which is then checked against a number of sources of guidance. Such practice

sounds rigorous and reliable, however, it also raises questions of manageability which could undermine the whole approach, since it would not be possible to collect this much evidence for every child, for every part of the curriculum. There is also an overwhelming focus on paper evidence which could endanger validity for those whose ability to write is not in keeping with their scientific reasoning. In Extract 6.4 the summative assessment appears to take place at the end of year, when **levels** are assigned. At this time the National Curriculum (DfEE 1999) levelling system was still in place, so it is perhaps surprising that the SL listed such a range of resources for the checking of judgements. The levelling system was still very much embedded in lesson planning during the DBR Phase 2 school visits, for example, on end of unit assessment record sheets (**B23-Ph2**). The SL explained how she used levelled outcomes to support her assessments within the lesson:

Extract 6.5

Focused LO and success criteria, levelled outcomes (no numbers next to children's names since carry plan around to annotate during lesson)

Biggest thing is that I know what I am looking for. Take concrete statements from QCA (eg Y5 L4 or Y6 L5) and then work out what comes before or after to create must/should/coulds. Use combination of other schemes and level descriptors to help decide this.

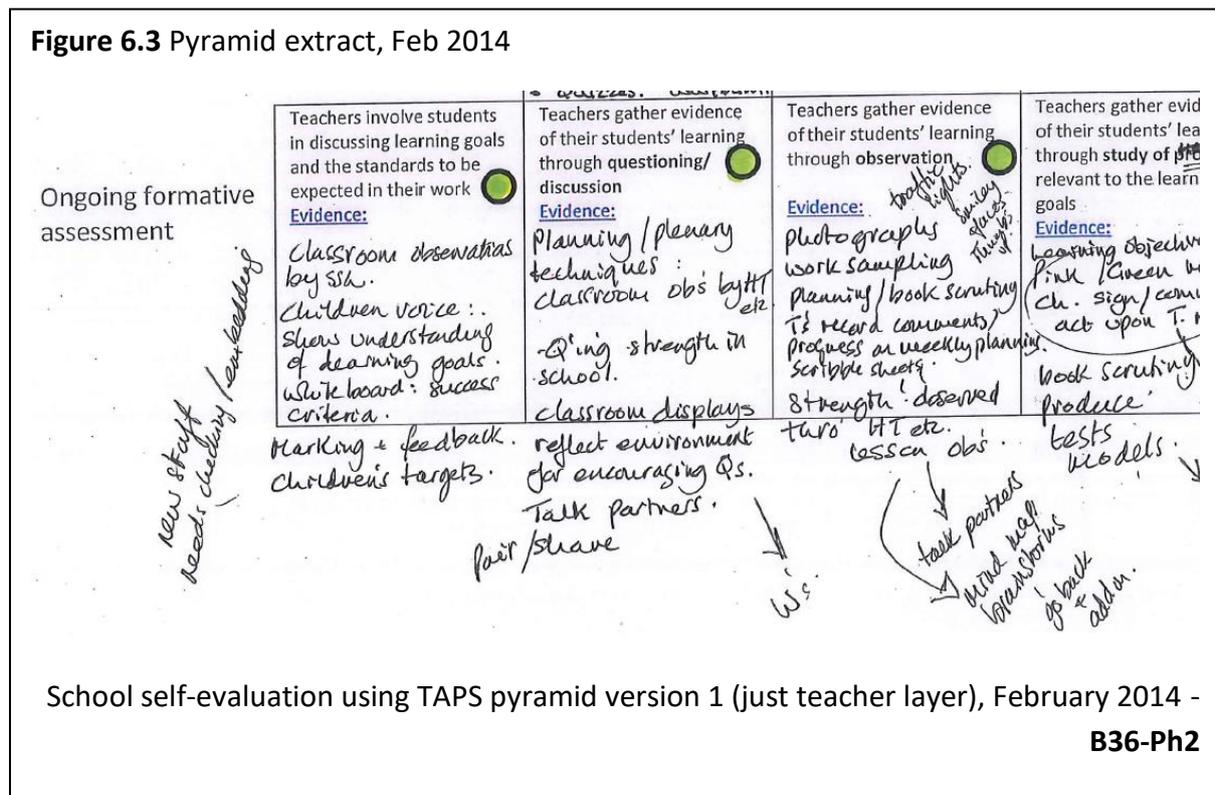
New curriculum with 2 year cycle could be more simple.

Y5/6 lesson observation discussion field notes, February 2014 - **B30-Ph2**

The use of criteria within the lesson will be discussed in Section 6.5 when considering practice at the Teacher layer; the extract is included here in order to explore the timing of summative assessments. Extract 6.4 suggested levels were used for summative end of term assessments, but examination of the teacher planning and discussion with teachers suggests that judgements were made against the levelling criteria during the sequence of lessons. The SL asserts in Extract 6.5 that the levels are not shared with pupils during the lesson, they are used to support teachers to: *'know what I am looking for'*. This provides a range of occasions to collect levelled outcomes, supporting the validity of summative assessments which summarise this range. However, it could have a negative impact on formative

purposes, with repeated levelling akin to frequent summative testing, which has been found to be detrimental to learning (Black 2012).

As noted in Section 6.2.2, the use of formative assessment information for multiple purposes needs careful guidance. However, the TAPS pyramid 'final product' and any associated guidance had not been produced at this time; so School B, as a TAPS project school, was using partly formed solutions which could have led to unpredicted consequences, as will be discussed further in Chapter 8. School B's emphasis on evidencing judgements could have been reinforced by the development of the TAPS pyramid self-evaluation tool which was introduced to project schools at Cluster Day 2, shortly after School B visit 2 in February 2014. This first version of the TAPS pyramid, asked schools to note evidence for their RAG (red, amber, green) rating and School B's notes were very detailed, as can be seen in the section presented in Figure 6.3.



Apart from talk partners and questioning, the majority of formative assessment strategies listed are paper-based or observed through lesson observations or work scrutiny. It could be argued that the TAPS pyramid appears to support a drive for the gathering of evidence.

The TAPS pyramid's emphasis on evidence, both for teachers gathering evidence of pupil learning and self-evaluation evidence of the strategies being used, could have influenced the school's examples to be physical recordings. Evidence collection and recording continued to be an area of focus, for example in the moderation staff meeting (Extract 6.1) and in the next steps identified for assessment in the second PSQM submission: *"Continue to develop our evidence collecting and AFL strategies to establish robust, confident procedures"* (March 2015, **B55-Ph3**).

The 'formative to summative' model within the TAPS pyramid suggests that information gathered during formative assessments can be summarised for summative purposes. However, a criticism of the model could be that it misrepresents formative assessment: it could appear that the main aim for classroom formative assessment is to supply information that can be used for summatively. If the formative is done for summative purposes, it effectively makes all assessment summative; formative strategies are only used as a way of getting to summative. This is similar to when Wiliam (2012) noted that AfL was misinterpreted by many to mean frequent summative testing. This indicates that the formative purpose represented in the TAPS pyramid model may need to be strengthened, especially when the model is enacted within a school system with high levels of accountability.

6.3.2 DBR Phase 3S - Confidence in teacher judgement

In DBR Phases 1 and 2, the SL listed a range of **structures** to support summative assessment, and an emphasis on **records** and **evidence**. In DBR Phase 3 there is more of a recognition for the role of the teacher:

Extract 6.6

Comment related to whether the TAPS project has increased understanding of role of assessment:

Given confidence to trust own opinion – given breadth of resources to validate this and recognize our (Teachers') judgements are valid.

Hearing a child is valid.

SL questionnaire, TAPS cluster day 7, November 2015 - **B78-Ph3**

The teacher is given a central role, with their '*opinion*' and '*judgements*' described as '*valid*'.

Confidence is a recurring theme in this Phase, with 11 out of its 13 coded occurrences in DBR Phase 3. It could be questioned whether a teacher's '*opinion*' would provide a reliable assessment, with a major criticism of teacher assessment being its potential for bias (Johnson 2012). However, the SL indicated that the teacher judgements are supported by a '*breadth of resources*', suggesting that the use of supportive structures remains integral.

The comment: "*Hearing a child is valid*" suggests that a previous emphasis on written evidence had not taken sufficient account of verbal interactions. This can also be seen in Extract 6.7 where the SL describes how they are no longer feeling that they need to write down: "*reams of what the children were saying*" to provide evidence.

Extract 6.7

It's become easier because it's not, "At the end of the unit, I need to look back through the books," or, "I need to look at photos," or, "I need to think about what so-and-so said." We are gathering that evidence as we go, making those judgments. To begin with, we might have written down reams of what the children were saying, but what do you do with that? You put it in a folder and then look at it again.

*We've found more efficient ways of doing that and trusting our judgment, each of those judgements, and saying, "I heard him say that. I can't possibly write it all down," because you can't, "but I knew what I was listening for, and I'm satisfied that that child said and did whatever it was that was required to match that. I've just ticked it. That's what was working for...he got it, he got it, he got it, but I am going to listen to her a bit more carefully because I'm not sure about her, so I'm going to focus on her this lesson." It's directed our teaching through our planning much more in a focused way in order to know what I'm looking for, am I finding it. I am with those children, and I need to * (16:50—unclear) some of those children. That hasn't become onerous; it has become upskilled, I'd say.*

SL interview, June 2016 - **B83-Ph3**

The SL suggested that their collection of evidence became more manageable, for example, rather than writing down everything which the children were saying in class discussions, the teachers were focused on what they were looking for in the lesson. She described the teachers as '*upskilled*' and the teaching and planning as more focused, indicating an increase in teacher assessment literacy (Klenowski and Wyatt-Smith 2014). There appeared to be less emphasis on recording and evidencing, with teachers making judgements in the lesson, rather than trying to prove them afterwards. Whilst this is described as more manageable by the school, it could raise questions in terms of reliability and validity, with teachers relying perhaps on their experience of the child's attainment in previous lessons or alternative subjects. Assumptions about pre-determined attainment will be explored further in Section 6.5.

6.3.3 Summary of changes at summative reporting layer

Key features of changes in assessment practice in this layer:

- Initially a large number of resources/structures/criteria lists were called upon to 'standardise' summative assessment, indicating a concern for consistency which will be explored further in the next section.
- Largely criterion-referenced assessment was in place: levelling criteria for assessments were used in lesson planning and at the end of term.
- A variety of methods (particularly paper based) were used to gather evidence of learning, seemingly for a primarily summative focus, raising concerns regarding interpretation of the TAPS pyramid as a model for collecting evidence for summative assessment, rather than a model with a strong formative base.
- DBR Phase 3 saw a move to more confidence in teacher judgements, including within lesson dialogue, with less emphasis on recording and evidencing.

6.4 Monitoring layer (M)

6.4.1 DBR Phase 1M - Concern for consistency

In DBR Phase 1 summative assessment appeared almost synonymous with **levelling**. Staff development, in the form of supporting documentation or staff meetings, was focused on gaining confidence with levelling, as described in Extract 6.8.

Extract 6.8

Teachers have been using a range of strategies for summative assessment of levels, including QC criteria, Rising Stars materials, Kent trust - and others which give useful indicators that I presented at a staff meeting and are listed on our server and copies in a file I presented to staff. Teachers now assign a level at the end of each unit taught.

PSQM C2 reflection, March 2013 - **B1-Ph1**

During the period of the case study the National Curriculum levelling system was removed and in 2013 teachers knew this change would be coming, which perhaps led to a search for possible alternative support with criteria. The SL describes an increasing list of supportive

structures (*QC criteria, Rising Stars materials, Kent trust - and others*) to support summative assessments.

The increasing list of supportive structures raises questions of manageability for staff to use all of these collected resources, together with possible issues with reliability if there are differences between the criteria for each. **Consistency** was noted by the SL as a key issue at the first TAPS cluster day (October 2013) and in interview the SL commented that: “*core principles for assessment in science have been established but there is difference in practice amongst classes*” (November 2013, **B10-Ph1**). Concerns regarding ‘consistency’ as an issue was predominantly coded in DBR Phase 1 (12 out of 15 occurrences), suggesting it became less of a concern later in the case record. There could have been a change in the meaning of consistency for the school over the case study period: at the beginning a consistent assessment approach appeared to mean doing the same, whilst Harrison and Howard (2009) argue that it is: ‘*consistency of principle not uniformity of practice*’, thus explicitly principled assessment could enable a range of practice to be ‘consistent’ (Davies et al. 2014). It is not clear from this data how the school or SL viewed the relationship between consistency and reliability; this wider issue will be discussed in Chapter 7.

6.4.2 DBR Phase 2M - Levelling and moderation

One of the reasons cited for the removal of levels was that it led to the labelling of children who began to speak of themselves as a ‘level 4c’ (Boaler 2015). The SL was keen to stress in the post lesson discussion that although she was using levels to guide her expectations for what she was looking for in the lesson, the level numbers were not shared with the children. The SL also spent time in a **moderation** staff meeting discussing the differences between assigning a level to a piece of work and assigning a level to a child:

Extract 6.9

What's required to level a piece of work?

- 1. Useful to see planning.*
- 2. Useful to know the context.*
- 3. Need to know what level/type of support might have affected outcomes.*
- 4. Useful to capture verbal comments during sessions.*
- 5. Useful to observe contributions and learning in sessions.*
- 6. Clear assessment criteria. Agreed sources.*
- 7. Good knowledge of curriculum content.*
- 8. Good knowledge of level descriptors.*
- 9. Good understanding of progression in skills and knowledge.*

What's required to level a child?

- 1. Evidence – as above – from a greater range of examples.*
- 2. Development can be seen over time.*
- 3. Teacher's records show progress.*
- 4. Listening to children 'talk science' and gauging breadth of thinking skills.*
- 5. Does a child's interest in a subject make a difference?*

Moderation staff meeting agenda for discussion, June 2014 - **B43-Ph2**

The distinction between work and child made here by the SL, could be aligned to the two different conceptualisations of summative assessment which were noted in Chapter 5: making a snapshot judgement of narrow attainment by levelling a piece of work from one context, or making a summary judgement of broader attainment by levelling a range of work from a child across numerous contexts. For the snapshot judgement, the SL emphasises the importance of 'context', suggesting that it is 'useful to see planning', 'know types of support' and 'observe contributions'. This in-depth knowledge of the context raises questions for the production of exemplification materials to support the implementation of the TAPS pyramid, which would be unable to capture this level of detail in a concise format. The SL also emphasises the importance of teacher knowledge and understanding of the subject 'progression', 'curriculum' and assessment indicators. This suggests a high level of pedagogical subject knowledge, which less experienced teachers may not have, raising questions regarding whether confidence in teacher assessment is a realistic expectation for

those who are newly qualified, since assessment competence needs a combination of skills (Black et al. 2011).

The summary judgement is more focused on *'development over time'* using *'a greater range of examples'*, which could support validity of assessment across contexts and instruments; more expansive rather than prescriptive assessment (Lum 2015). *'Evidence'* and *'teacher records'* are listed, but pupil talk and thinking skills are also mentioned, perhaps suggesting a widening of what is considered appropriate assessment information on which to form a judgement.

The moderation meeting provided dedicated time for the teachers to discuss how they were making their judgements (**B40-Ph2**). It was not a simple checking of levels assigned to individual pieces of work, the aim appeared to be more to develop the assessment literacy of the staff, to make explicit the tacit knowledge of how to make judgements (Sharpe 2004).

6.4.3 DBR Phase 3M - Range of information

In DBR Phase 3, the SL describes the supportive nature of moderation and *'sharing practice sessions'* in developing confidence in teacher judgements:

Extract 6.10

We have moved from 'each teacher doing their own thing' to having basic frameworks, resources, levelling references and expectations in place – but with the freedom for teachers to try what works within a framework of sharing and discussing with each other about what is being tried. Moderating sessions and sharing practice sessions keep the progress moving forward – with the aim of teachers being confident in their judgements because a) they know their curriculum expectations and b) they know what meeting them 'looks like'.

PSQM C2 reflection, March 2015 - **B55-Ph3**

There still appears to be a tension with regards to consistency, how much to stick to the *'framework'* and how much *'freedom for teachers to try what works'*, indicating a rigidity dimension to teacher conceptualisations of assessment. However, the *'sharing and discussing with each other about what is being tried'* suggests a process of reflection and

evaluation of strategies to support teachers to actively construct their practice (Sharpe 2004). Such dialogue could support development at both an individual and whole school level (Stoll et al. 2006).

In 2013 a large number of published materials were listed to provide structure or criteria in support of summative assessment (Extract 6.4, **B3-Ph1**). By 2016 the SL advocates a different approach: summative assessment which is not described separately, or based on separate materials, but is ongoing and informed by formative assessment (Extract 6.11).

Extract 6.11

Outcomes for our school [related to reliability]:

- *Understanding that useful, reliable assessment opportunities come from good, consistent and varied science teaching where opportunities to assess against the requirements are frequent.*
- *Assessment criteria is built into the planning stage – with learning objectives and success criteria made explicit to the children.*
- *Massive reduction in reliance on or requests for summative testing materials / papers to validate, confirm or substitute for teacher’s judgements.*
- *Much discussion about what assessment is – several members of staff participated in TAPS sessions. All sessions fed-back at staff meetings.*

SL presentation planning, May16 - **B80-Ph3**

The SL suggests that ‘consistent’, ‘reliable’ judgements have been supported by including the assessment criteria at ‘the planning stage’, assessment is part of teaching and this enables ‘frequent’ assessment opportunities. This appears to enhance validity, with multiple and ‘varied’ assessments able to capture a broader range of the curriculum than is possible in end of term snapshots, but the question remains as to whether the frequent assessments are detrimental to the formative purpose. Enacted practice of ‘formative to summative’ assessment will be explored further in the next section.

6.4.4 Summary of changes at monitoring layer

Key features of changes in assessment practice in this layer:

- An early concern with consistency and standardisation of practices, which led to cross-checking with multiple structures. A later tension between ‘sticking to the framework’ and allowing more ‘freedom’ indicated a rigidity dimension to teacher conceptualisations of assessment.
- Moderation discussions supported teacher assessment literacy and indicated two contrasting conceptualisations of summative assessment, in line with findings in Chapter 5, which were enacted as: levelling of work (snapshot) and assigning a level to a pupil (summary).
- There was an ongoing attempt to balance validity and reliability, with recognition of the importance of a range of information for valid assessment, whilst development of reliability through use of assessment criteria throughout teaching and learning; supporting a shared understanding and assessment literacy.

6.5 Teacher layer (T)

6.5.1 DBR Phase 1T – Strategies include marking

A range of **strategies** to elicit and record pupil ideas were represented in the case record. One recurring theme was a teacher focus on **marking** and children responding to marking, perhaps in response to the schools’ Ofsted (2012b) report recommendation to improve this noted in Section 6.1.2:

Extract 6.12

Elicitation and assessments such as floorbooks, concept cartoons are regularly used. Marking is ‘pink and greens’ throughout with a clarification type comment and next steps identified.

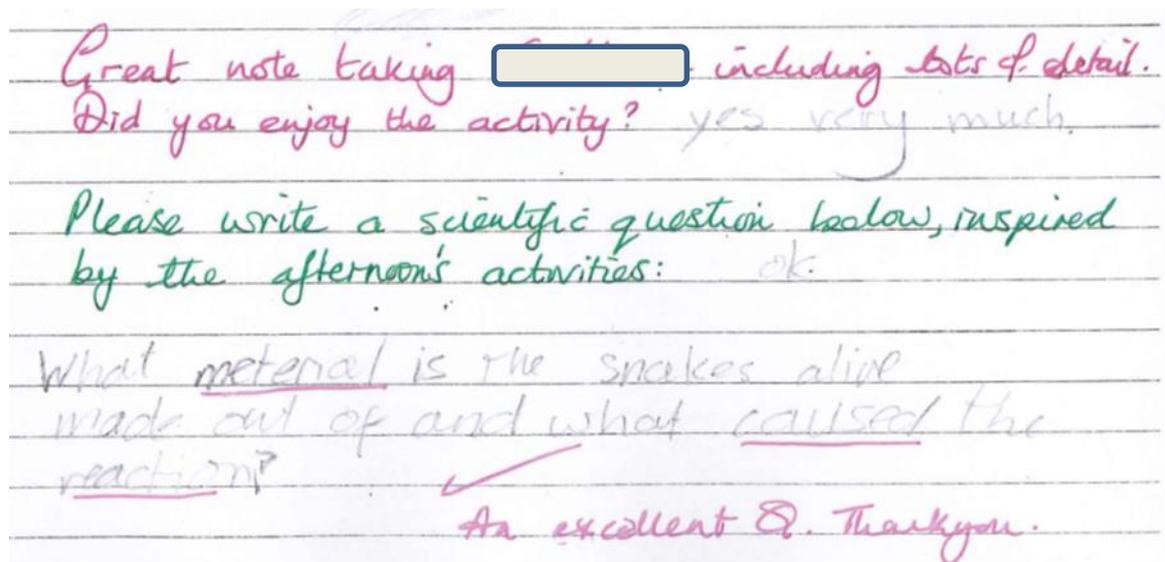
SL interview field notes, November 2013 - **B10-Ph1**

‘Pink and green’ marking refers to the school’s marking policy at the time which was to use a pink pen for positive feedback (‘tickled pink’) and a green pen to provide next steps (‘green for growth’). Such marking was seen in work samples collected at the time (**B15-Ph1**) and

on subsequent school visits. The practice of using a range of colours when marking was popular around this time but has more recently been questioned due to the increased workload as teachers write a large number of comments, and then have to re-mark once the child has responded (Marking Policy Review Group 2016). An example of such 'triple' marking is shown in figure 6.4:

Figure 6.4 Marking example

Pink and green marking after recording of an indoor fireworks demonstration:



Y5/6 pupil work samples provided, collected November 2013 - **B14-Ph1**

The work samples were provided by the SL, so are likely to represent what the SL sees as best rather than typical practice; such triple marking may have been more typical for marking of English writing. Nevertheless, when visiting the school in February 2014, field notes include comments about the extensive marking in the Y3/4 class: "Detailed marking of previous work shows 'tickled pink' and 'green to grow'" (B22-Ph2). The detailed marking raises questions about manageability, together with the value placed on written recording – a particular issue for younger children which will be discussed further in Section 6.6 when considering self-assessment. In contrast to School A, the policy is for 'comment only' marking (Butler 1988), promoting task-involvement rather than ego-involvement (Harrison and Howard 2009). Nevertheless, it appears to be the role of the teacher which dominates,

the teacher is the one with the pink and green pens, marking after the lesson, the one designating the next step, again an area for further discussion when considering the role of the pupil in Section 6.6.

The marking attempts to play a formative role with its next step comments and the resulting responses from pupils, but it could also be playing a summative role as the teacher works through the books making judgements regarding attainment within the lesson. The pink marking provides recognition of attainment, a summative judgement, whilst the green marking provides the formative next step. However, it could also be argued that the pink marking provides the pupil with examples of their success, supporting an understanding of the features of a 'good' piece of work. The written dialogue between pupil and teacher could represent a formative scaffolding to support the pupil to develop their ideas. Nevertheless, this is all based on the assumption that the pupil reads the marking and is active in their response. With such a time implication for teachers to triple mark and a delay to the feedback to pupils, the consequential validity of such a recommendation may be compromised, with triple marking being unsustainable.

6.5.2 DBR Phase 2T - Making assessment manageable

School B explored a range of strategies over the case study period in the attempt to find workable, manageable solutions. One of these was to narrow the focus for teacher attention by pre-defining success criteria (levelled planning, **B29-Ph2**) and expectations for pupil outcomes in the form of **differentiation**. For example, in the Y5/6 lesson on mixing materials observed in February 2014, pupils worked in pre-designated attainment groups and were given recording sheets with varying amounts of scaffolding. This differentiation featured in the pre and post-lesson discussions with the teacher:

Extract 6.13

Discussion with teacher before lesson:

Seated in maths groups, but tweaked and may mix.

Know who are: must / could / should.

Annotate planning during or after lesson, then hand in planning

Know what looking for and know children because already grouped them

Lesson observation notes regarding grouping:

'By end of lesson...' -success criteria which are on the plan

Must, should, could expectations for task

Sitting in ability groups with targeted activities. Some have scaffolded sheet where need to circle, middle have gaps sheet, others write from board. TA with one group.

Y5/6 Lesson observation field notes, February 2014 - **B30-Ph2**

The recording sheets provided a different amount of structuring and challenge for different groups, with for example: the simplest sheet directing the children to identify if a new product was made, whilst the more complex sheet asked the children to explain (**B32, B33-Ph2**). Such structuring or scaffolding represents a balance for teachers: enough structuring so that pupils are successful, but not so much that the task becomes too easy; judging and adapting this level of challenge within the lesson is one of the features of Assessment for Learning (Loughland and Kilpatrick 2013). However, by setting the challenges before the lesson, the teacher had pre-decided the pupil outcomes. This appears particularly clear upon examination of the pre-prepared end of lesson expectation grid (Figure 6.5), which had pupil names already typed underneath:

Figure 6.5 End of lesson expectation grid

2014 Term 3 'Over reactions'

LA:

Use an ipad magnifier to support increased detail in on drawings and descriptions...

MA:

...Use equipment effectively to improve detail in observations and make links between them...

HA:

...Use the evidence obtained, linked to other knowledge and experience to develop a hypothesis...

2014 T3W5 Over-reactions!		
LA:	MA:	HA:
<ul style="list-style-type: none"> Use an iped magnifier to support increased detail in drawings and descriptions. Work with others to make decisions about scientific ideas. Explain that in an irreversible change, the materials cannot return to their original form. That they have been changed forever. 	<ul style="list-style-type: none"> make generalisations, based on evidence and experience. Use equipment effectively to improve detail in observations and make links between them. Explain the main criteria that would determine if a change was irreversible and give examples of materials and processes that would demonstrate this. (include gas) 	<ul style="list-style-type: none"> Give reasons why this collaborative approach to experiments may improve the validity of evidence but recognise its limitations too. Use the evidence obtained, linked to other knowledge and experience to develop a hypothesis. Consider social issues related to the advancements in changing materials. Explain how the knowledge of the difference between physical and chemical changes helps explain irreversibility, with examples! (molecules)
<p>So We ---</p>	<p>H Ti O</p>	<p>Ma Jb i Cf</p>

(Names under each column denoting Low/Middle/High Ability)

Y5/6 Lesson observation plan, February 2014 - **B29-Ph2**

The teacher had grouped the children on the assessment record, to make the task of observing in the lesson more manageable, so that she 'knows what she is looking for' as she described in Extract 6.13. The SL may be clear about what pupil outcomes to look for in the lesson, an example of convergent assessment (Torrance and Prior 1998), but she has also decided what she expects each child to attain: the children's names were already typed underneath the success criteria lists ready to be ticked. The children were also seated in attainment groups and given varying amounts of scaffolding or TA support. It could be argued that the lesson was a checking of the teacher's assumptions based on the children's prior attainment. Pupil outcomes had already been closed down, with attainment being capped for some, for example, the group who work with the Teaching Assistant did not have the chance to show that they could work independently or make decisions (Boaler 2015). In order to make teacher assessment within a practical lesson manageable, the SL appeared to narrow the task to evidence gathering in support of a pre-existing summative assessment. The information could be used formatively, since those who did not perform as expected

could be given further support in the following lesson, but the emphasis appeared to be on confirming summative judgements.

Formative assessment can include adapting subsequent tasks to match pupil needs, for example, by providing support with access or differentiated levels of challenge in activities. However for summative purposes, if different support or tasks are given, there could potentially be doubts raised regarding whether a reliable comparison could be made. This has implications for 'formative to summative' assessment since it suggests that reliable summative assessment may not be able to be based on differentiated activities. In addition, pre-determined task and group allocation could effectively put a ceiling on attainment for some (Boaler 2015). Also, if tasks are too closed, the outcomes also become predictable for the pupils, such predictability in assessment can be a form of construct irrelevance, with 'rote learned responses' rather than a 'demonstration of skills' (Stobart 2009: 168) and 'box-ticking' rather than in-depth learning (Mansell et al 2009).

Later in the year, at the moderation staff meeting, one of the concerns raised was: *"How much support is too much? [The] dilemma is how much is scaffolding, what support has been given"* (June 2016, **B40-Ph2**). This raises a question of whether the activity should be unsupported if an assessment is to be used summatively; in Vygotskian terms this is assessing 'actual development' rather than identifying the 'zone of proximal development' (1976). This raises questions for the use of formative assessment information being used for summative purposes, since if using an activity for both purposes, there is the implication that it should be unsupported, leaving little space for the teaching. The scaffolding role of the teacher is integral to teaching and learning (Tharp and Gallimore 1988), and if it needs to be withdrawn for an assessment to take place, then we are back at the frequent testing criticism. Development in Vygotsky's 'proximal' or 'potential' zone involves carefully structured interventions (Alexander 2008), like those under the banner of Assessment for Learning. Such formative assessment is an ongoing process, which is perhaps why School B spent so much energy trying to capture them. The discussion, interaction and learning, develops in a continuous manner; it is perhaps in the pauses, the teacher's questions in mini-plenaries or the time for self-reflection where the fruitful assessment opportunities arise. The final chapters will discuss the contention that not all formative assessment may

be suitable for summative purposes, since if the focus is purely on independent ‘actual’ development, then opportunities for learning may be missed.

6.5.3 DBR Phase 3T - More open

There is some evidence that differentiation became less closed, with children choosing their level of challenge for homework activities (Spring 2015, **B66-Ph3**) and that grouping became more mixed rather than by prior attainment (May 2016, B81-Ph3). The SL also mentioned the rise of open ended inquiry in the final interview, suggesting this was an ongoing area of development:

Extract 6.14

*The impact has been—in order to access all the areas of the assessment, we are ensuring that we are giving opportunities for children to demonstrate theirs. Particularly at the top end of the school, where there’s greater independence required, they have to demonstrate that they have, with understanding, selected this system or selected that piece of information or present it in that way. We’ve got to give them those open-ended inquiry tasks where they can demonstrate that it has been their choice, not that * (5:45—unclear) told them that. Open-ended inquiry is definitely launching as well.*

Transcribed SL interview, June 2016 - **B83-Ph3**

The SL describes how pupils are given more ‘opportunities’ to demonstrate their understanding and independence, particularly via ‘open-ended inquiry’, a process in which they are engaged and active. The opening out of activities could provide for a wider range of pupil outcomes; more divergent assessment (Torrance and Prior 1998). When considering the implications for a ‘formative to summative’ model of assessment, there could be a balance between providing open and closed tasks. If the task is too open, then the outcomes may be useful for formative assessment but not for summative judgements against criteria. Whilst if the outcomes are too closed, with tightly defined success criteria, then the task becomes a tick-list: more reliable to judge, but less valid in its sampling of the curriculum. Perhaps here is the division between open elicitation, of the kind which can be done at the beginning of the unit to gather information about what children already know; and more focused activities, where there is a clear learning objective or success criteria for judgement. The former open tasks are useful for formative assessment, but the latter could

be useful both formatively and summatively. The focused activities need to be open in the sense that a range of outcomes could be produced, and this is not pre-decided by the work which is given. The open-closed continuum is clearly a dimension of assessment which impacts on enactment of assessment practice.

Whilst School B appeared to have moved to a more open approach in terms of the setting of tasks, there is little in Extract 6.14 to support discussion of the role of group work. Social constructivist theories are relevant when considering whether summative assessment of an individual's attainment assessment should be based on collaborative endeavours. As discussed above (Section 6.5.2), collaborative group work and teacher scaffolding require further guidance in 'formative to summative' assessment. It would appear simplest to say that only individual work can be used for summative assessment, but this then excludes the large amount of group practical work done in primary science, together with downgrading the value of social interaction in learning (Rodrigues 2004). Perhaps the school's emphasis on evidence noted earlier represented an attempt to find a way to make individual assessments from classroom practices which involved working in groups. These issues will be revisited in Chapter 7 when considering guidance for practitioners regarding the types of tasks which could be used for 'formative to summative' assessment.

6.5.4 Summary of changes at teacher layer

Key features of changes in assessment practice in this layer:

- An early emphasis on feedback through detailed marking raised questions regarding manageability and impact.
- Attempts to make whole class assessment more manageable by pre-defining groupings and outcomes led to closed convergent tasks based on teacher assumptions.
- It was suggested that not all formative assessment may be suitable for summative assessment of independent outcomes, for example, due to the amount of teacher support provided. This has implications for a 'formative to summative' model of assessment.
- DBR Phase 3 saw the beginning of moves towards more open divergent tasks, allowing for a range of pupil outcomes. It was noted that the open/closed continuum was a dimension of assessment for which teacher guidance could be generated from this study.

6.6 Pupil layer (P)

6.6.1 DBR Phase 1P - Trialling strategies

During DBR Phase 1, the SL lists a range of **strategies** for assessment:

Extract 6.15

Post-it notes are used effectively at F/Y1/Y2 - where a teacher/TA annotates what children have said in plenaries/elicitation exercises or discussion with the teacher or peers. This is then attached to the relevant learning evidence. One teacher annotates a paper copy of her lesson plan with comments children made at each stage, which works effectively for her. At KS2, children are more involved with self-evaluation of the lesson achievement and use various devices to record this: eg, smiley faces, today I have learnt....statements.

C2 reflection, PSQM submission March 2013 - **B1-Ph1**

In some comments there appears to be a concern to use the strategies as a way of providing or gathering **evidence** of children's ideas, with little mention of the formative purpose or use of the evidence. The focus on evidence could indicate the strategies being used to gather information for summative assessment, as discussed in Section 6.3. There appears to be a sense of trialling strategies to find out 'what works' for different teachers.

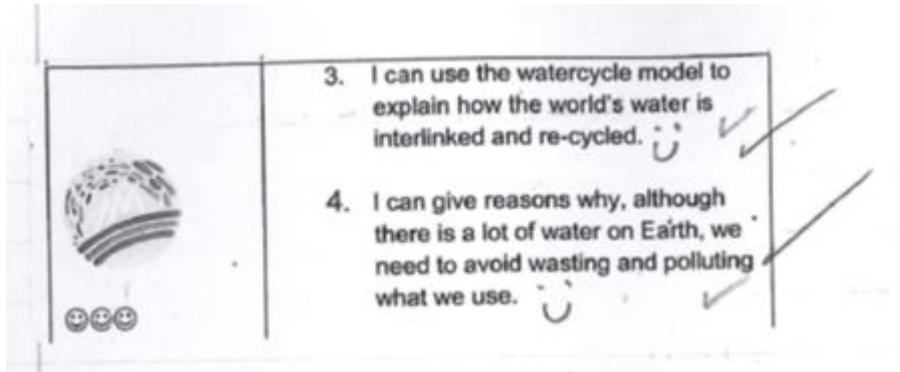
Pupil self-assessment is noted at Key Stage 2, recorded by writing statements about what they have learnt or using smiley faces, which could signify confidence levels with particular areas of learning. The SL describes KS2 children as '*more involved with self-evaluation*'. It is unclear from this extract why this is limited to children in KS2; perhaps it is due to the emphasis on recording strategies, or it could be merely that the SL, as a KS2 teacher, knows more about developments in Year 3-6.

It is unclear whether the self-assessments are used formatively or summatively, although the recording strategies appear to suggest a judgement is being made about attainment in the lesson: '*today I have learnt...*'. Examples of the self-evaluation stickers were seen in

children's work collected during this time (Figure 6.6), together with pupil responses to marking, which were already discussed in Section 6.5.1.

Figure 6.6 Self-evaluation sticker at end of topic on water cycle

3. *I can use the watercycle model to explain how the world's water is interlinked and re-cycled.*
4. *I can give reasons why, although there is a lot of water on Earth, we need to avoid wasting and polluting what we use.*



Y5/6 pupil work samples provided, collected November 2013 - **B14-Ph1**

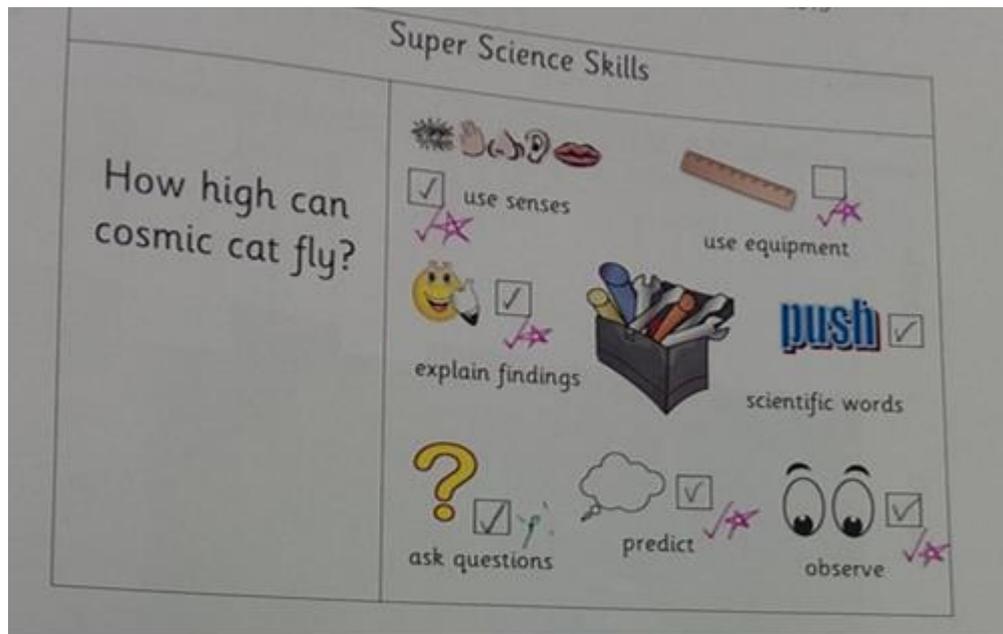
On the self-evaluation sticker, the child has drawn smiley faces and a tick to show that they think they have achieved the objectives; the teacher has also ticked to show that they agree. The stickers, and the triple marking discussed above, require the pupil and teacher to look again at their work, with the teacher in a checking or confirmation role. The **next step** comes from the teacher, raising questions about how actively the pupils are engaged in the self-assessment.

6.6.2 DBR Phase 2P – Developing strategies for self and peer assessment

School B graded themselves as confident green for all elements of the TAPS pyramid pupil layer except for **self and peer assessment (B53-Ph2)**. On the self-evaluation pyramid, the teachers noted a 'science tool kit' which was in development, which referred to a pictorial tick sheet created by the Y1 teacher, pictured in Figure 6.7.

Figure 6.7 Key Stage 1 science toolkit

Super Science Skills for 'How high can cosmic cat fly' investigation: use senses, use equipment, explain findings, scientific words, ask questions, predict, observe



Science toolkit for KS1, example collected March 2015 - **B67-Ph3**

In their design of a Key Stage 1 'toolkit', the school sought to make self-assessment accessible to young children by including pictures and minimal recording. In this example, the pupil and teacher appear to have arrived at slightly different judgements about which science skills had been developed during that investigation. It is interesting to consider issues of validity arising from this, for example, whether the teacher has a better understanding of skills progression and therefore would be able to make a more reliable judgement, but if the child knows they have not understood the focus on using equipment in this lesson then their judgement could be more valid. This raises questions regarding aggregation of pupil and teacher assessments, if pupil self-assessment is to be included in summaries of attainment, and adds the dimension of 'assessor' to conceptualisations of assessment.

During the DBR Phase 2 staff meeting described above (Extract 6.1), the self-assessment labels (**B48-Ph2**) were described as: "*redundant because still had to assess work*". This

suggests that the pupil's self-assessment was not considered useful at this time because it was the teacher who needed to assess the work, to make the judgements, perhaps returning the pupil to a more passive role.

6.6.3 DBR Phases 3P - Developing the role of the pupil

Turning to consideration of how the SL felt the school's use of formative strategies had changed during the DBR Phase 3:

Extract 6.16

Staff are actively trying, evaluating and sharing strategies. For example, we have always done elicitation exercises at the beginning of units of work – but formats were repetitive and little was done with them.

Our Y1 and Y2 teachers worked together to produce a new format that enabled their children to add reflective comments at the end of a unit for comparison.

The Y3+4 teachers have 'Wonder Walls' where children post-it note questions for science related to the new topic and as answers are discovered and discussed – they are grouped separately.

Y5+6 have used a quiz for instant responses – with True/False/? (no idea) options – then questions created and used again at the end of the unit and discussion time given.

Children understand the purpose of doing this and are proud of what they have learned.

Teachers use a variety of additional strategies for formative assessment, including: observations, self/peer assessments, stickers for instant feedback specifically created for a target and note-taking on planning as a lesson is delivered.

PSQM C2 reflection, March 2015 - **B55-Ph3**

The exploration of different strategies for different age groups has continued, but there appears to be a recognition at the beginning of the extract that elicitation should lead somewhere, that something should be done with the information gathered (line 2-3), although the SL does not explain here what is done with the information. The primary purpose for the strategies appears to be formative, although there are two suggestions (lines 5 and 9) that the children make an 'end of unit comparison', signifying that the pupils may be involved in reviewing their progress.

The school trialled a range of strategies, but rather than feeling they needed to adopt them all (which is an area noted for summative assessment structures in section 6.3), they have chosen the ones which matched their topic or year group. The emphasis on development of peer/self-assessment was explained in more detail in a presentation by the SL:

Extract 6.17

Traditionally we have for a long time done thought showers, concept maps, at the beginning of a unit of work and we realised we were just rolling that out as something to do because that's what you are supposed to do, without necessarily understanding it or measuring its impact or using it in any valid way, so we started then to come back to those thought showers at the end of a topic and annotate them and add on to that to see the progress and the difference. And then we just branched out completely and we've looked at lots of different strategies for self-assessment and peer assessment on the basis that if you get children involved in what they're doing and understand the process you're putting them through, they'll become better at it because they'll understand where you're taking them and they'll go there with confidence and understanding but if you are just doing it to them then they don't feel that either they want to put anything into that process and they don't get as much out of it.

SL presentation, June 2016 (transcribed from video) - **B85-Ph3**

The SL describes a shift from elicitation activities which are 'done to' the pupils as part of the 'traditional' planning, to actively involving the pupils in the process. The SL's description resonates with the way William (2011) describes how formative assessment can 'activate' the learners, enhancing their learning. Such metacognition may feed into pupil summative summaries of their learning, but no mechanism for this is described here. The SL does go on to describe a change in the role of the teacher, as more 'power' is placed in the hands of the pupils:

Extract 6.18

This one here [pointing to photo] is where we use pupils now, we're developing their strategies as both assessors and ambassadors and as mentors. So this one is an image of lower able children working together trying to develop those skills of communication, problem solving and setting up an investigation with a [pupil] mentor who's just ace at it, who's there to support and to lead. Not to bossy and not to tell how, but to suggest how about, before you do that how about thinking. So she's leading, she's me doing that job, probably better than me because they'll take it from her more than from me.

SL presentation, June 2016 (transcribed from video) - **B85-Ph3**

The changing role for the pupils perhaps signifies a change in teacher role, which the SL noted when explaining how their use of peer assessment had developed and the pupils began to 'do her job'. The pupil 'mentor' is acting as the more knowledgeable other (Vygotsky 1978), supporting their learning in the ZPD. It is unclear whether such experiences enhance summative assessment opportunities, but they may broaden the range of formative assessment data available for summary.

6.6.4 Summary of changes at pupil layer

Key features of changes in assessment practice in this layer:

- Many strategies were trialled and evaluated throughout the case study period, initially with a focus on evidence gathering for summative purposes, later on involving younger learners.
- In DBR Phase 3 a greater emphasis on formative assessment was demonstrated, with recognition that strategies are only formative if the information is utilised, suggesting the formative purpose was no longer subsumed by a drive for summative evidence.
- Self and peer assessment were key areas of development for the school, developing from teacher-led strategies, to attempts to actively involve learners. In line with Case Study A, pupil-led assessment was used for primarily formative purposes, adding an 'assessor' dimension to conceptualisations of assessment.

6.7 Conclusion

6.7.1 Key features of changing assessment practice at School B

In this analysis the following key features of changes in science assessment at School B have emerged, organised by pyramid layers (in bold):

- **Whole school processes:** early conceptualisations of assessment included dimensions of timing, value judgements, formality and separation. Summative purpose and processes appeared to take priority during DBR Phase 2. Later conceptualisations indicated a closer relationship between formative and summative assessment, a focus to *'embed AfL'*, with the continuous gathering of data from formative assessment to inform the 'summative grade' so that summative assessment was no longer a separate *'bolt on'*. This indicates a move towards a 'formative to summative' model, but it is important to note that this was not until the third year of the case study.
- **Summative reporting layer:** criterion-referenced summative assessment initially called upon a large number of structures to support standardisation. There was an increasingly strong emphasis on gathering evidence, raising concerns regarding interpretation of the TAPS pyramid as a model for collecting evidence for summative assessment, rather than a model with a strong formative base. DBR Phase 3 saw the beginning of a move towards more confidence in teacher judgements (of their own judgements and on the part of the subject leader), with less reliance on recording every utterance and cross-checking with multiple structures or criteria lists.
- **Monitoring layer:** initially there was a concern for consistency, later there was recognition of the importance of a range of information, but with shared assessment criteria and consideration of a 'rigidity' dimension in the conceptualisation of assessment. Moderation discussions were used for levelling and developing teacher assessment literacy, and included the distinction between the two conceptualisations of summative assessment noted in Chapter 5: snapshot (when looking at work samples) and summary (when considering summative assessment for a child).
- **Teacher layer:** there was an early emphasis on feedback through detailed marking which raised questions regarding manageability and impact. Attempts to make whole class assessment more manageable by pre-defining groupings and outcomes led to closed convergent tasks based on teacher assumptions. Moves towards more open tasks were seen in DBR Phase 3, allowing a range of pupil outcomes. The open/closed continuum was identified as a dimension for teacher guidance, acknowledging that not all formative assessment may be suitable for use in a summative summary.
- **Pupil layer:** a range of strategies were trialled, initially with a focus on evidence gathering, later considering ways to actively involve learners through peer and self-assessment. This added an 'assessor' dimension to conceptualisations of assessment, and indicated a greater emphasis on formative purposes by the end of the case study period.

This three year case study provides an exploration of changing practices over time. With such a long time span, it is impossible to address every issue within the school, so this chapter focused on changes pertinent to the relationship between formative and summative assessment. It appeared that the school was trying to use a ‘formative to summative’ approach to assessment, in line with the Nuffield (2012) and TAPS approaches. However, this took a long time to develop, and during DBR Phase 2 it appeared that involvement in the TAPS project was increasing the emphasis on evidence and summative purpose, subverting the formative purpose; an issue which will be revisited in Chapter 7.

6.7.2 Tentative generalisations

As noted in Chapter 5, tentative or ‘fuzzy’ generalisations (Bassey 1999) acknowledge the uniqueness of the context, whilst also noting features which could have relevance for other contexts. Tentative generalisations which can be drawn from the case study of School B suggest:

- Teacher conceptualisations of formative and summative assessment may include a range of dimensions: timing, value judgements, formality, separation, rigidity, assessor and open/closed.
- Pre-determined pupil outcomes via grouping or differentiated recording may reduce validity of assessments. Activities that are open provide opportunities for range of outcomes, but needs to be focused on particular objectives to allow for both formative and summative uses.
- Teachers need to trial strategies to make them ‘work’ for their context. Change in assessment practice takes a substantial amount of time for it is intricately entwined with teaching and learning.
- Summative assessment can be snapshot or summary judgements; it may be the latter which can be informed by formative assessment information.
- The TAPS pyramid could be interpreted to mean all assessment opportunities should lead to information which can be used for summative purposes, subverting the formative purpose. If ‘formative to summative’ assessment is to support pupil learning, then clear guidance regarding formative purposes is needed.

Such tentative generalisations identified through the PSQM, School A and School B data, provide new insights into the conceptualisation and enactment of the relationship between formative and summative assessment in primary science. Chapter 7 will draw together and discuss the findings from the three data chapters.

Chapter 7 Discussion

7.1 Introduction

This study sought to develop understanding of the relationship between formative and summative teacher assessment of primary science in a sample of schools in England, using a Design-Based Research (DBR) approach to develop guidance for practice. A theoretical model of ‘formative to summative’ assessment, proposed by the Nuffield expert group (2012) and operationalised by the TAPS project (Davies et al. 2014), was used as an analytical framework to support a comprehensive analysis of data from the Primary Science Quality Mark (PSQM) and two case studies which explored the relationship between formative and summative assessment in action.

The chapter will begin with a brief summary of key findings in response to the research questions (RQs):

RQ1. How do teachers assess children’s learning in science for **formative and summative** purposes?

RQ2. How can teachers’ conceptualisation and enactment of the **relationship between formative and summative assessment** of children’s learning in science be used to inform guidance for practice?

RQ3. How can study of **changes over time** in conceptualisation and enactment of the relationship between formative and summative assessment be used to inform guidance for practice?

The discussion will then explore the following three areas arising from the study:

- Conceptualisation and enactment of the relationship between formative and summative assessment across the datasets (Section 7.3).
- Validity and reliability in ‘formative to summative’ assessment (Section 7.4).
- The relationship between formative and summative assessment represented in the TAPS pyramid ‘formative to summative’ model (Section 7.5).

This chapter will draw together findings from Chapters 4 to 6 to explore the three areas for discussion, and propose new theoretical and practical products to support the relationship between formative and summative assessment in primary science.

7.2 Summary of key findings in response to research questions

7.2.1 Formative and summative assessment in primary science (RQ1)

In response to RQ1, a wide range of formative and summative assessment strategies were catalogued in the PSQM dataset and both case studies. In terms of formative assessment, the PSQM schools listed a range of elicitation strategies which included: teacher-led talk, collaborative activities, observation and paper/task-based activities, but it could not be assumed that these strategies always fulfilled the formative purpose of informing teaching and learning. In School B (DBR Phase 1) an early emphasis on feedback through detailed marking raised questions regarding manageability and impact (Section 6.5.1). By DBR Phase 3 a greater emphasis on formative assessment was demonstrated in School B, with the school trialling more open, divergent activities for scientific inquiry (Torrance and Prior 1998) (Section 6.5.3).

In terms of summative assessment, nearly all of the PSQM schools described at least one method of assessment which could be categorised as summative, with tests and tracking grids the most frequent methods, but there was little explanation of the processes involved (Section 4.2). Over a third of the PSQM schools (37%) described using a combination of summative methods, but the process of aggregating data was not explicit. In School A, summative assessment was conceptualised as a summary, enacted as a 'best fit' judgement which aimed to draw on a range of information (Section 5.3.1), but which may also have masked gaps in attainment. There was no set process for making a 'best fit' judgement and it was suggested that such judgements may rely on in-depth teacher knowledge of the subject; which may make it difficult for inexperienced teachers, without further guidance and exemplification.

7.2.2 Relationship between formative and summative assessment (RQ2)

In response to RQ2, the lack of a relationship between formative and summative assessment was a key finding in much of the data. Formative and summative assessment were largely described separately in the PSQM dataset, with some schools also describing completely separate systems for conceptual understanding and inquiry skills (Section 4.2). School A also separated summative judgements for concepts and inquiry skills (Section 5.3.2), raising questions for how science was broken down atomistically and then recombined into a holistic judgement. Formative and summative assessment appeared to be separated on a value dimension, with summative assessment described more negatively. Other dimensions, marking differences in the way formative and summative assessment were perceived included: timing, degree of formality, separation from typical classroom activities, rigidity, teacher or pupil as assessor and how open or closed the activities were. A framework for these dimensions will be proposed in Section 7.3.

The relationship between formative and summative assessment was supported by shared school structures and moderation discussions. For example, School A utilised school-wide progression structures for both formative and summative criterion-referenced assessment which impacted on staff planning, science coverage and assessment confidence.

Moderation discussions at School B made the process of making summative judgements explicit and indicated two contrasting conceptualisations of summative assessment which were enacted as: levelling of work (snapshot) and pupils (summary) (Section 6.4). Making such a distinction explicit in a 'formative to summative' approach will be discussed in Section 7.5.

7.2.3 Change over time (RQ3)

In response to RQ3, the three year case study of School B identified an ongoing attempt to balance validity, reliability and manageability. For example, early concerns for consistency and standardisation of practices led to cross-checking with multiple structures in an attempt to strengthen teacher assessment reliability. Later, attempts to make whole class assessment more manageable by pre-defining groupings and outcomes led to closed convergent tasks based on teacher assumptions (Section 6.5.2). DBR Phase 3 saw the

beginning of moves towards more open divergent tasks, allowing for a range of pupil outcomes (Section 6.5.3) and more valid teacher assessment. Study of such changes over time indicates the need for explicit recognition of such a 'balancing act', which will be discussed further in Section 7.4.

An additional finding from the three year case study of School B was that for some time there was an increasingly strong focus on gathering evidence for summative assessment, raising concerns as to whether the 'formative to summative' approach was conceptualised as repeated summative assessment rather than a summary of formative assessment. The TAPS pyramid could be interpreted to mean all assessment opportunities should lead to information which can be used for summative purposes, subverting the formative purpose. If 'formative to summative' assessment is to support pupil learning over time, then clearer guidance regarding formative purpose and 'formative to summative' processes is needed, which will be discussed in Section 7.5.

7.3 Conceptualisation and enactment of the relationship between formative and summative assessment

7.3.1 Dimensions in the conceptualisation and enactment of the relationship between formative and summative assessment

A number of dimensions have been identified throughout the study which can provide insight into the way teachers conceptualise and enact assessment practices. A summary of the dimensions is presented in Table 7.1, in order to map the way teachers in the sample understood formative and summative assessment.

Table 7.1 Dimensions in the teachers’ conceptualisation and enactment of assessment

	Dimension	Associated with formative	Associated with summative
Outcome	Purpose	Support learning, to inform planning	Accountability, to provide a number for tracking
	Value	‘Good’	‘Bad’
	Audience	Pupils/teacher	Teachers/senior leaders/external
	Intention	Divergent (what they know)	Convergent (whether they know x)
	Evidence	Could be ephemeral or ‘in the teacher’s head’	Largely paper based
Framework	Reference	Ipsative (pupil)/criterion referenced	Norm/criterion referenced
	Scale	Short term (lesson) goals More atomistic	Longer term goals More holistic
	Assessor	Pupils/teacher	Teacher/external
	Timing	Frequent, within lessons	End of term/topic, one off
	Separation	Ongoing, part of teaching, typical classroom activity	Snapshot, special, ‘bolt on’ activity or Summary/‘best fit’
Classroom practice	Strategies	Open strategies	Closed strategies
	Formality	Informal	Formal
	Rigidity	Flexible	‘Stick to plan’ for consistency
	Support	Pupil can be supported	Pupil must be independent

In much of the data analysis a separation was found between formative and summative assessment, with teachers providing different descriptions along a number of dimensions. There is no suggestion that every teacher will conceptualise assessment along all of these dimensions, but a number of these may be present. To support the discussion below, each dimension is in bold type.

Value judgements regarding assessment were a recurring theme in the data (e.g. Sections 5.2 and 6.2). Formative assessment was seen in a more positive light, with its **purpose** to support learning and plan for next steps, whilst summative assessment was described more negatively and often associated with accountability and the collection of paper-based **evidence** for an external **audience**. Harlen (2013) asserts that summative assessment’s ‘poor reputation’ stems from the dominance of measured performance, which has left little time for formative assessment practices (p23). A redefinition of summative assessment in terms of the use of its summary role may be needed for teachers, to enhance the

relationship between formative and summative assessment; this will be discussed further in Section 7.3.2.

The dominant form of assessment **reference** for both formative and summative judgements was found to be criterion-referenced assessment (e.g. Sections 5.4, 5.5.2 and 6.4).

Nevertheless, teachers did use norm-referencing during moderations discussions, where comparative judgements appeared to support the development of a shared understanding about progression (e.g. Sections 5.4.2 and 6.4.2). Ipsative-referencing was utilised when discussing individual pupil progress (e.g. Sections 5.6 and 6.6), but this appeared not to be utilised in the same way, with ipsative comments useful in summaries for parents, whilst criterion-referencing was needed to track performance.

The enactment of criterion-referenced assessment is largely dependent on the criterion **scale** in use. This was particularly pertinent during a time of curriculum change, for example, School B appeared to be searching for more detailed criterion lists (Section 6.4.1) and School A continued with their criterion scales from the previous curriculum (Section 5.4.1). It appears that the previous National Curriculum levelling system criteria (1999) were so broad, as they were designed for a 'best-fit' approach and not for fine-grained measures of progress (e.g. sub-levels), that schools felt they needed to translate them into more manageable steps, for example Science Stars (School A, Section 5.5.2) or child-friendly 'I-can statements' (PSQM data, Section 4.3). The new National Curriculum (2013) of criterion-referenced Age Related Expectations (ARE) contains finer grained conceptual objectives (although the Working Scientifically objectives are still broad), which do not need to be translated in the same way, perhaps providing a stronger base for a shared understanding. The ARE provide more atomistic lesson objectives and the programme of study is put forward as the new 'Attainment Targets' (DfE 2013a), whilst the levelling system could arguably be described as more holistic. However, the ARE criteria still need to be aggregated or summarised in some way to provide an overall summative assessment, thus the move from atomistic to holistic judgements remains an issue.

The role of the pupil in the assessment process was an area of development for many schools (PSQM Section 4.3 and School B Section 6.6) with initiatives to try to involve pupils

more actively in self and peer assessment. It was found in the case study schools that the pupil as **assessor** was largely utilised for formative purposes, to support learning within the lessons. Pupil assessments did not appear to feed into summative assessments, perhaps because teachers did not trust the pupils to make accurate assessments, or perhaps because they felt that the pupils did not have the broader understanding of the curriculum which would enable their assessments to be accurate: they did not share the same understanding of progression as the teachers. It is unclear from the study of practice in these schools, whether the pupil role should develop further to become more active in summative assessment processes, or whether pupil assessors would be most suited to formative assessment only. If summative assessments were more ipsative, then pupil assessments of their own progress would be a valuable contribution, but whilst summative assessment is criterion-referenced, the broader knowledge of the curriculum criteria is held by the teacher rather than the pupil, making teacher assessment the primary source of summative judgements.

Teachers in the sample separated formative and summative assessment in terms of **timing**, with the latter designated by occurring at the end of a period (e.g. Sections 4.2, 5.2 and 6.2). It was also often seen to be separate, a 'bolt on' (Section 6.2.3). This **separation** from normal classroom activities is important because it makes summative assessment appear to be something special, something separate from everyday teaching and learning, which is problematic for a system of 'formative to summative' assessment, indicating that such a system would require changes in conceptualisations as well as practice. The view of summative assessment as a snapshot or a summary is an important distinction for a 'formative to summative' model and will be discussed further in Section 7.3.2.

The **formality** of an assessment appeared to define whether the assessment was used for formative or summative purposes, with the latter being more formal or uniform (Black et al. 2011), with clear 'rules' to be followed and a **rigid** plan (Section 6.5). A key dimension for these 'rules' is how much **support** to provide for the children. It is assumed that summative assessment recognises independent achievement, the 'fruits' rather than the 'buds' of attainment (Vygotsky 1978). This represents a real conceptual issue for a 'formative to summative' model because the teacher will provide a range of support within classroom

activities, leading to rich formative assessment data, but it could be questioned whether such data could be used to inform summative assessments if varying levels of support were provided. There is also the potential for a negative 'backwash' (Isaacs et al. 2013) from summative assessments if it is assumed that only independent data can be utilised, and so teachers feel that they need to 'stand back' within class rather than support the learning process. The formative purpose will be 'driven out' if the summative purpose becomes dominant in the classroom (Harlen 2013: 23). Such an effect was arguably seen for a time at School B when an emphasis on evidence and consistency was prevalent (Section 6.3.1). An important implication is that any model of 'formative to summative' assessment must accentuate the formative purpose so that this does not get subsumed by the summative drive.

The **intention** of assessment could be divergent or convergent, to find out what the learner knows, or whether the learner knows something (Torrance and Prior 1998). The use of divergent, open **strategies** to explore the learner's ideas were more associated with formative assessment or ipsative progress over time, for example, by returning to mindmaps or KWL grids (PSQM data, Section 4.3). Such open activities are difficult to summarise for summative purposes because they are not closely tied to the criterion scale, for which more convergent, closed strategies would be more suitable. For a 'formative to summative' model this means that not all types of activities would be equally useful for summative purposes, suggesting a need for guidelines regarding the kinds of strategies which could be utilised (which will be the focus of Section 7.3.3).

Each dimension discussed above (and summarised in Table 7.1) was presented as pairs of extremes, if formative and summative assessment were viewed completely separately; however, this study seeks to explore the relationship between them, with a view to providing recommendations for a 'formative to summative' approach. Thus Table 7.2 identifies a new way of looking at these dimensions which reconciles the 'extremes' by providing a pathway between them, developing links between formative and summative assessment. Depending on the context and purpose of different assessments, teachers may prioritise different dimensions to optimise their approach.

Table 7.2 Dimensions in the relationship between formative and summative assessment

	Dimension	Primarily formative	A 'formative to summative' approach	Primarily summative
Outcome	Purpose	Support learning, to inform planning	To support and summarise learning	Accountability, to provide a number for tracking
	Value	'Good'	All assessment can have positive and negative consequences	'Bad'
	Audience	Pupils/teacher	Pupils/teachers/senior leaders/external	Teachers/senior leaders/external
	Intention	Divergent (what they know)	Opportunities for divergent and convergent	Convergent (whether they know x)
	Evidence	Could be ephemeral or 'in the teacher's head'	A range of evidence is valued	Largely paper based
Framework	Reference	Ipsative (pupil)/criterion referenced	Largely criterion referenced	Norm/criterion referenced
	Scale	Short term (lesson) goals More atomistic	Focused goals linked to curricular objectives	Longer term goals More holistic
	Assessor	Pupils/teacher	Largely teacher-led	Teacher/external
	Timing	Frequent, within lessons	Ongoing, with periodic summaries	End of term/topic, one off
	Separation	Ongoing, part of teaching, typical classroom activity	Range of activities, with periodic summaries	Snapshot, special, 'bolt on' activity or Summary/'best fit'
Classroom practice	Strategies	Open strategies	Range, including focused assessment strategies	Closed strategies
	Formality	Informal	Range of activities	Formal
	Rigidity	Flexible	Flexible within boundaries of objectives	'Stick to plan' for consistency
	Support	Pupil can be supported	Level of support taken into account in judgements	Pupil must be independent

A 'formative to summative' approach requires a clearer relationship between formative and summative assessment, which is the aim of the 'middle ground' of Table 7.2. The middle column contains a range of proposals to support the development of a 'formative to summative' approach which seeks to balance concerns of validity and reliability; thus providing the basis for discussion in the ensuing sections. In Table 7.2 it is noted that both formative and summative **purposes** have **value**, in an attempt to counter the perception that summative assessment is inherently 'bad'. In order to strengthen the validity of assessments, it is suggested that **evidence** should draw on a range of activities in terms of: **intention, formality** and **strategies** (to be explored further in Section 7.3.3). By drawing on a range of activities, the **timing** of assessment is more ongoing, feeding into a summary of attainment, rather than basing a summative judgement on a single **separate** snapshot (which will be discussed next, in Section 7.3.2). In order to strengthen the reliability of assessments, it is suggested that there is a **rigid** focus on criterion-**referenced** objectives, with the level of **support** taken into account in teacher judgements. There is also a recognition that not all formative assessment needs to inform summative judgements, the teacher **assessor** is best placed to make attainment summary judgements because they have access to both the **scale** of the curricular objectives, together with the ongoing and wide-ranging evidence. Nevertheless, consideration of so many dimensions draws attention to the complexity of the process and the demands it will place on the teachers' assessment literacy, an area for focus in Section 7.4.

It is important to note that a 'summary' conceptualisation of summative assessment is embedded within the 'formative to summative' column in Table 7.2, which requires more detailed exploration in the next section. It is suggested that this view of summative assessment provides a link between formative and summative assessment, where the latter is part of the teaching and learning process, rather than a separate 'bolt on' (Section 6.2.3).

7.3.2 A distinction to support the relationship between formative and summative assessment: summative assessment as snapshot or summary

An emergent finding from, in particular Case Study B, was that summative assessment was conceptualised in two different ways: as an attainment snapshot or summary (Section 6.4.2). The focus for the snapshot or summary in this discussion is criterion-referenced pupil attainment, rather than a summary of ipsative progress or evidence. School B's moderation staff meeting (June 2014, B43-Ph2) contained discussion of the difference between: 'levelling a piece of work', where attainment in a stand-alone activity provides information about a snapshot in time; and 'levelling a child', where attainment in a range of activities across the term or year is summarised. Both of these are summative assessments involving grading or levelling and judging against criteria, but they are quite different in terms of the information feeding in and the inferences which can be made from them.

An attainment snapshot could be a sample of work or the result of an end of a unit test or task, which provides information about attainment in a particular context at a particular time. Such snapshot assessments may be completed in more formal or standard conditions, supporting reliable comparison, but inferences should be limited to attainment regarding that small part of the content domain (Stobart 2009). In addition, snapshot assessments could be seen to be quite separate from classroom teaching, supporting a separation and polarisation in the conceptualisation dimensions of formative and summative assessment. In contrast, an attainment summary judgement needs to consider a number of activities, taking into account their context and timing, which could enhance construct validity by drawing on a wider range of information, some of which may have been collected for primarily formative purposes. Thus for a summary judgement, summative assessment can be informed by and arise from both formative assessment and summative snapshots, in line with the 'formative to summative' model proposed by Nuffield (2012) and TAPS (Davies et al.2014). Formative and summative assessment can be viewed separately if summative assessment constitutes only snapshots, but if the summative assessment is to be a summary of attainment, then the 'formative to summative' model can be applied.

Lum (2015) has noted a recent paradigm shift, from summative assessment as a way of sorting individuals, to summative assessment as a description of what individuals know and

are able to do. Snapshot assessments can provide information for sorting or comparing individuals, perhaps in standard conditions, whilst summary assessments provide a description of an individual over time and across a wider range of contexts, providing a larger sample of the content domain.

School A's use of 'best fit' judgements were akin to summary assessments, but the 'best fit' was related to the wide ranging level statements of the 1999 National Curriculum, where a level could be assigned, which may have masked gaps in understanding. Initial guidance for assessments in the newer curriculum, the 'Interim Teacher Assessment Framework' (STA 2015), was designed to lead to a more 'mastery' approach, where all elements were required before the pupil could be said to be meeting expectations. However, concerns were raised by schools feeling either that they needed to collect written evidence for each objective (as was seen in School B), or that such a 'secure fit' judgement pertaining to all statements was unmanageable for teachers (DfE 2017), leading to updated guidance moving to a 'best fit' for writing in English and an instruction for science that schools should only have demonstrable evidence for content taught in the final year of the key stage (STA 2017). Schools where summative assessment is seen as separate to normal classroom activities, as special snapshots, sat in special conditions, understandably would feel that evidencing the whole curriculum would be unmanageable. However, if summative assessment is seen as a summary, drawing on a range of activities, then a 'secure fit' judgement seems more feasible. Nevertheless, this requires subject, pedagogical and assessment literacy on the part of teachers. A shared understanding of progression in the subject is required to be able to make judgements within a lesson which can be summarised at a later date.

7.3.3 Strategies for teacher assessment

A wide range of assessment strategies were catalogued through the study, which could be utilised for formative or summative purposes. Nevertheless, in the PSQM submissions there were some strategies which were found to be more aligned with one purpose, for example, using a KWL grid formatively. Drawing together the findings from the PSQM database (Chapter 4) and the two case studies (Chapters 5 and 6), assessment strategies can be classified into three categories: open, focused and closed (Table 7.3), with the middle

‘focused’ category particularly supporting the ‘formative to summative’ process described in Table 7.2.

Table 7.3 Examples of different types of teacher assessment strategies

Open elicitation assessment strategies which could be higher on validity but lower on reliability <i>(good for eliciting ideas at the start of a topic to inform planning)</i>	Focused assessment strategies which could balance validity and reliability demands <i>(good for tracking progress of inquiry skills across the year, to inform summative summaries)</i>	Closed snapshot assessment strategies which could be higher on reliability but lower on validity <i>(good for a quick check of concepts – summative snapshots which can inform summative summaries)</i>
Pupil question raising KWL grid (Know, Would like to know, Learnt) Mind map/thought shower Pupil drawing Concept cartoon discussion Open investigations Pupil presentation Observation of pupil explorations	Focused teacher questioning Whole investigations with focused recording of one element Choice of challenge tasks Observation of pupil explorations supported by a Working Scientifically tracking grid or expectations on planning Feedback or marking focused on the objective Self/peer assessment using success criteria	Written tests Quick fire teacher questions Multiple choice quiz Cloze-the-gap or matching activities Diagrams with pre-made labels A directed sorting activity (only 1 way to sort)

Table 7.3 does not provide an exhaustive list of assessment strategies, nor does it guarantee validity or reliability, but it provides a starting point for practitioners to consider how to balance open, divergent strategies, with convergent, closed strategies (Torrance and Prior 1998). Embedded within the table are a number of the other dimensions of assessment conceptualisations which were discussed in Section 7.3.1. For example, the dimension of time is included to draw attention to open strategies providing useful information to inform planning at the beginning of the topic, whilst focused assessment approaches can aid the tracking of progress through the year. The value dimension is present in the identification of recommendations for what the strategies might be ‘good for’, addressing the concern that one type of assessment is inherently ‘bad’ (Harlen 2013).

The validity and reliability of each strategy would depend on the way it is used in class, the criteria it is based upon and the way teachers interpret responses, for example, teacher questioning could switch between open exploration of children's ideas to quick fire closed questions. Nevertheless, some strategies are likely to provide pieces of information which are useful for reliable comparison between pupils or classes, whilst other strategies could provide more valid or authentic information about what the child is able to do, but which may be hard to record or compare.

A degree of divergence in tasks may be necessary to assess inquiry skills in action, indicating that different types of strategies may be useful for teacher assessment of inquiry skills and conceptual understanding. Between open and closed assessment, a third category of focused assessment is listed in Table 7.3, with the claim that such strategies could support the tracking of progress across the year, particularly for inquiry skills. For example, children can carry out a full investigation but only a part of this is the focus for pupil recording and teacher assessment, as recommended by McMahon and Davies (2003) and operationalised in the TAPS focused assessments. Such focused assessment would be criterion-referenced, allowing a judgement against curricular objectives which could feed into a summative summary, but it should also allow for diversity of outcomes so that the task is not too narrow. Focused assessment is different to a summative snapshot, because for a snapshot the whole task is the assessment, whilst for a focused task, the assessment is a part of a larger task. For example, the pupils may be engaged in a whole investigation rolling cars down ramps, but the focus for assessment could be on the recording of results. Focused assessment provides a bridge between formative and summative, but guidance and examples will be required to support teachers to implement such an approach.

7.3.4 Summary of conceptualisation and enactment of the relationship between formative and summative assessment

Key ideas arising from this section:

- A range of dimensions in the conceptualisation and enactment of teacher assessment were identified during the study including: value, timing, formality and

support. Such dimensions provide a framework for discussion of teacher assessment and also indicated the level of challenge in changing practice since this may require a number of the dimensions to be addressed.

- One key dimension was the conceptualisation of summative assessment as snapshot or attainment summary, the latter was identified as the one which would result from a ‘formative to summative’ model of assessment.
- Teacher assessment strategies from the study data were categorised to provide examples for teachers of open/divergent, focused and closed/convergent assessment, all of which could be used for any purpose, but with focused assessment identified as the most likely to provide information which could be used both formatively and summatively.

7.4 Validity and reliability in ‘formative to summative’ assessment

7.4.1 Discussion of validity and reliability in ‘formative to summative’ assessment

One reason for implementing a ‘formative to summative’ approach is that it would mean summative assessments could draw on a richer base of data, providing a more valid sampling of the subject. Black and Wiliam (1996) argued that a comprehensive picture of achievements could be built up by aggregating different assessments which had been designed for a formative purpose. However, Gardner et al. (2010) assert that at the heart of all assessment should be a concern for improving learning and if this is subsumed by a summative focus then the assessment process will have little validity. A key concern for ‘formative to summative’ assessment is that the purpose of the assessment does not become confused, serving neither the formative nor summative purpose well.

Such a confusion of purposes could be seen in School B’s ongoing concern for consistency and evidence, prioritising reliability, perhaps at the cost of validity. The focus on evidence arguably led to repeated summative judgements (Section 6.2.2), a concern to ‘level’ or judge at each interaction. This represents a misinterpretation of a ‘formative to summative’ approach: that all assessment was ultimately for summative purposes. Taras (2005) argues that all assessment is summative, but whilst all assessment does involve a judgement, in

comparison to peers or criteria, it is not accepted that all assessment needs to lead to evidence for later summaries. The formative purpose within a 'formative to summative' approach needs to be strengthened to ensure that summative evidence gathering does not dominate.

Alternatively, the emphasis on evidence seen in DBR Phase 2 at School B could be interpreted as a positive move in terms of reliability, with teachers basing their judgements on evidence rather than assumptions. Gipps et al. (1995) found improvements at the introduction of statutory teacher assessment, where practices moved from an intuitive approach to one based on evidence and written records (p176). Therefore, some emphasis on evidence to support reliable judgements is necessary, but it was found in School B that over-emphasis on evidence may have a negative effect on formative purpose. Davis (1998) suggested that strictly limited tasks would not provide evidence of 'rich knowledge', goals should be learning focused rather than performance focused. The goals and criteria need to be transparent and clear to both pupils and teachers, they: "*ought to share a particular conception of the knowledge and skills which pupils are supposed to be acquiring*" (Davis 1998: 152), indicating a role for moderation discussions, which will be discussed further below.

DBR Phase 3 for School B appeared to mark a shift in thinking from the concern for evidence and reliability, to consideration for validity and the role of the pupil. DeLuca et al. (2016) suggest that teacher assessment literacy should be reconceptualised as a developmental process. When the Subject Leader (SL) commented that: '*hearing a child is valid*' for example (Extract 6.6, B78-Ph3), it suggested development in teacher assessment literacy: a broadening in understanding of the types of information which can be used for assessment, which could lead to a wider sampling of the curriculum. For the teacher to know what to do next after 'hearing the child' is dependent on a certain level of understanding on the teacher's part though, for without an understanding of progression within the subject then it would be difficult to make a judgement or decide a next step; to move from 'evidence gatherers' to 'systematic planners' (Gipps et al. 1995).

Black et al. (2011) suggest that there needs to be a balance between uniformity and diversity in practice, since some degree of uniformity is required to support moderation discussions. Uniformity also supports routine and controllable activities, which feel 'safer' for the non-specialist (Abrahams and Millar 2008), which links back to the rigidity and formality dimensions discussed in Section 7.3.1. However, uniform rigid practice may not respond to the needs of the pupils, damaging the formative processes. Perhaps there needs to be a clearer assertion that not all formative assessment can or should be used to also serve summative purposes (Black and Wiliam 1996).

An interesting feature of School A was the way that their structures for assessment (e.g. criterion lists like Science Stars) remained largely stable through a period of National Curriculum change and involvement with the TAPS project, an area of consideration with the DBR process (Section 8.2.2). The structures played an important role in developing and sustaining a shared understanding of progression, providing shared criteria for planning, teaching and assessment. The National Curriculum (DfE 2013a) objectives also aimed to provide shared criteria, removing a separate level descriptor for assessment (DfE 1999), however, the wide-ranging statements, particularly for Working Scientifically, have required clarification and exemplification (STA 2016), which is perhaps another reason why School A continued with their own structures, until further guidance was available.

School A's structures for assessment included a separate system for inquiry skills. 18% of the PSQM schools described using tests for assessment of conceptual understanding and a 'tracking system' for assessment of inquiry skills. This could represent the two different conceptualisations of summative assessment noted in Section 7.3.2: the snapshot and the summary, with schools using a snapshot assessment for concepts and a summary judgement for inquiry skills. This could be a response to the nature of the knowledge being assessed: with conceptual knowledge assessed at a point in time, whilst inquiry skills considered over a longer period. Such practice assumes that inquiry skills and conceptual understanding can be separated, whereas many researchers argue that inquiry skills are deeply embedded within a context (Millar 2010). Attempts to separate science into atomistic pockets could hamper scientific literacy, with concepts being seen as separate to scientific method.

Alternatively, by utilising both snapshot and summary systems it could be argued that a wider range of assessment data is being collected, sampling a broader range of objectives thereby enhancing validity. Thus snapshot assessments would record attainment in conceptual understanding at a particular point in time, whilst ongoing tracking of skills could feed into an attainment summary of inquiry, providing two separate assessment judgements for different features of science. Nevertheless, if the systems are run separately it could be both unmanageable and unhelpful for understanding of the nature of science and for assessment literacy, with a lack of clarity in the processes and purposes of assessment. However, it could be possible to utilise both conceptualisations of snapshot and summary within one system of assessment: within a 'formative to summative' model, attainment snapshots could feed into an overall summary judgement which could combine all elements of the primary science curriculum; an area for further discussion in Section 7.5.

Harlen (2007) asserts that teacher assessment can be as reliable as it needs to be with moderation. 21% of PSQM schools made comments about moderation, whilst moderation was coded 30 times for School A and 29 times for School B. Nevertheless, the meaning of 'moderation' could be contested, with some referring to a process whereby judgements were checked, with a concern for inter-rater reliability (Johnson 2013) and others referring to a process of professional dialogue where the meaning of criteria (Section 5.4.2) or types of evidence were explored (Section 6.4.2). Connelly et al. (2012) found that explicitly stated curricular descriptors provided a common language for the teachers to use in assessing pupil work, which in conjunction with moderation and exemplification, meant that teachers arrived at more consistent judgements. However, they also noted that new processes could initially challenge teacher confidence and 'their current status as experts' (Connelly et al. 2012). Exemplification of moderation for professional learning could be another DBR product from this study and the TAPS project (Earle and McMahon 2017).

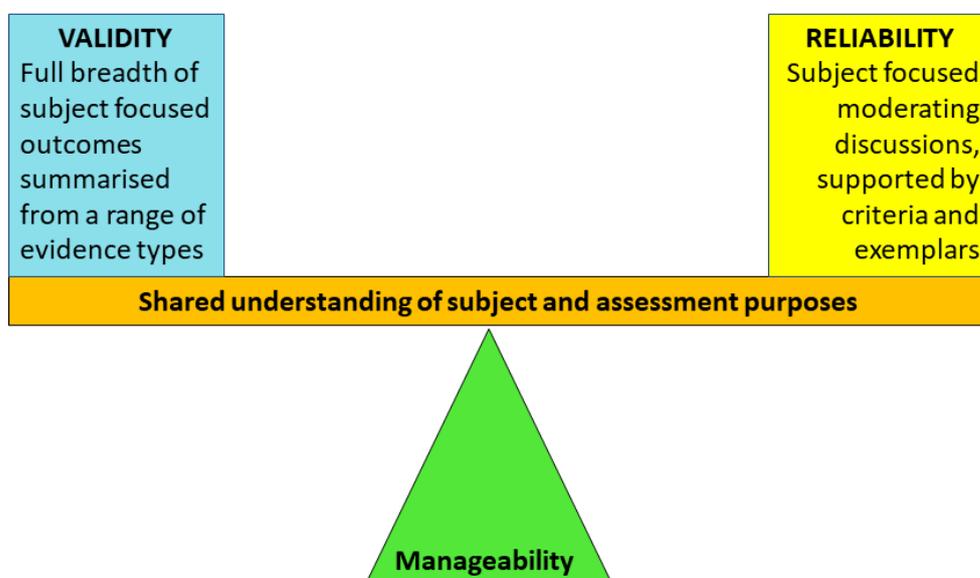
A shared understanding across the school, of science and of assessment, appeared to be enhanced by a criterion structure and moderation discussions (Sections 5.4 and 6.4). This shared understanding or shared criteria meant that formative assessment could be

summarised for summative purposes because both assessments were using the same benchmarks for decision making.

7.4.2 A seesaw model of teacher assessment

One of the difficulties for teacher assessment appears to be the balancing act between valid assessment of the whole of a detailed curriculum, and maintaining reliable, consistent judgements: Wiliam's (2003) 'trade off'. Harlen (2007: 23) states: '*an assessment cannot have both high validity and high reliability*'; it is not possible to have highly repeatable, standardised assessment which samples the whole of practical primary science. In School B, the development of assessment practices over time indicated an ongoing attempt to balance validity, reliability and manageability; whilst School A had an embedded system of assessment which was based on a shared understanding of progression in science. These features of the case study schools are represented in Figure 7.1 which provides a way of representing the balancing act; providing guidance to support validity by basing judgements on a broad range of information, whilst supporting reliability of judgements by utilising shared criteria, exemplars and moderation.

Figure 7.1 The Teacher Assessment Seesaw: balancing validity and reliability when using formative assessments for summative purposes.



In order for this representation to do more than merely describe the problem for teacher assessment, the detail within the ‘balance’ is designed to support teacher assessment literacy, providing guidance in principle rather than specifics, recognising the diversity of approaches required depending on the context of the assessment (DeLuca et al. 2016). The aim is to both develop teacher understanding of terms like validity and reliability, together with beginning to suggest principles for practice. Of course, translation of such a complex issue into a seesaw analogy diagram necessitates losing detailed meanings, such as the multi-faceted nature of validity. It is accepted that the Seesaw model is something of a simplification, but it is also proposed that such a simplification could support teacher assessment literacy by active engagement in discussion of the assessment principles of validity and reliability (DeLuca and Johnson 2017). Each feature of the model is discussed below, followed by exploration of two examples to further consider the ‘balance’.

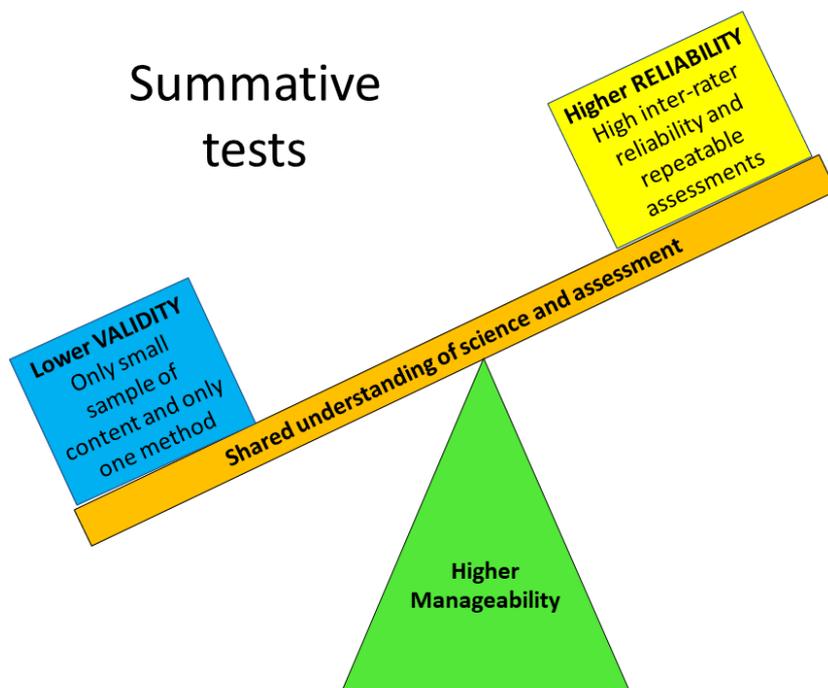
In this model:

- **Validity** focuses on content validity, equated with providing a summary of the child’s performance throughout the whole of the curriculum, which for primary science includes scientific inquiry, to combat *construct-under-representation*. The suggestion here is that any summative reporting should be based on a range of evidence types, which aims to reduce the *construct irrelevance* e.g. whether the child can read the question (Black and William 2012).
- **Reliability** is supported by reference to criteria (e.g. the Interim Teacher Assessment Framework STA 2015), exemplars (e.g. STA 2016 exemplification, TAPS database www.pstt.org.uk) and moderating discussions where teachers consider from different perspectives what it means for a child to have met a particular objective. Such moderation meetings with colleagues support teachers to be confident and more consistent in their judgements, but it is important that these discussions are focused on the science objectives to avoid unconscious bias from assumptions about the child’s behaviour or performance in other subjects (Campbell 2015).
- **Manageability** is explicitly highlighted at the base of the seesaw because if the ‘weight’ of number of assessments to satisfy both validity and reliability concerns are too onerous for the teacher, the manageability fulcrum will collapse.
- **Shared understanding** is the ‘beam’ on which the other concepts rest, since assessment literacy, together with a secure grasp of progression in the subject area, underpin teacher assessment. To be able to balance concerns of validity and

reliability, teachers require an understanding of what these terms mean for their context, what constitutes valid assessment and the criteria by which reliable judgements are made. The school community should work towards a shared understanding of the nature of primary science, for example, by discussing their expectations for progression in science skills and concepts. There also needs to be a shared understanding of the purposes of assessment: that it can be primarily formative, to support pupil progress and that this can be summarised at different reporting points as necessary. If assessment is only understood in terms of testing, then it devalues inquiry skills which are not easily tested, and it removes the active involvement of pupils to direct their own learning (William 2011). Discussing formative and summative assessment, with reference to criteria and exemplar benchmarks, supports teachers to be confident and consistent in their judgements.

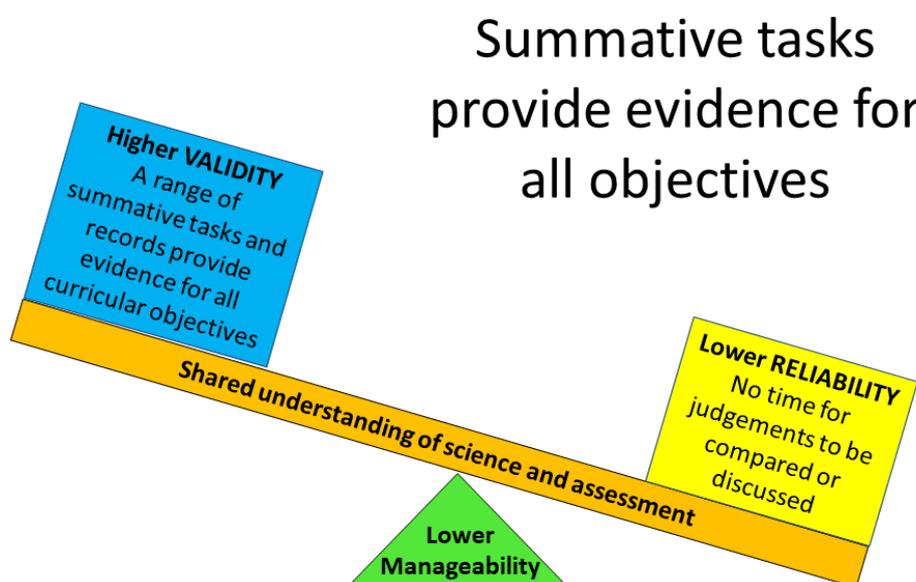
The Seesaw model can be further explained by brief consideration of two examples (Figure 7.2 and 7.3). Figure 7.2 displays the Seesaw balance for snapshot ‘summative tests’ which are represented as manageable to deliver and with high inter-rater reliability, but with lower validity for practical primary science because they can only sample a small part of the content domain. Such a concern was seen in both the literature (e.g. Mansell et al. 2009, Gardner et al. 2010) and the PSQM database, where schools were trying to use a combination of methods.

Figure 7.2 Teacher Assessment Seesaw for summative tests in primary science



Whilst Figure 7.3 displays the Seesaw balanced in the opposite direction by considering the gathering of written evidence from summative tasks for all curricular objectives in primary science. Such a method would arguably provide a more valid assessment of the content domain by providing a broader sample, but it would be unmanageable and leave little time for moderation to support reliability. This is akin to repeated summative assessment, rather than a summary of attainment as described in a ‘formative to summative’ model, since the purpose is primarily summative, repeated summative assessments to gather evidence, as seen for a time in School B.

Figure 7.3 Teacher Assessment Seesaw for summative tasks for all objectives



The proposed balance in Figure 7.1 encapsulates many of the dimensions from Table 7.2: embracing a range of outcomes and evidence types to support validity (related to dimensions of: intention, evidence, formality, strategies, timing and separation); whilst building a shared understanding to support reliability via moderation, criteria and exemplars (related to dimensions of: reference, rigidity, support and scale).

The Seesaw model (Figure 7.1) is designed to support teacher assessment literacy by identifying key concepts in assessment; including ‘social moderation’, where teachers

discuss and debate, to help build a shared understanding (Klenowski and Wyatt-Smith 2014). Brown (2004) notes that teachers need multi-dimensional models of assessment, a simple ‘summative bad’-‘formative good’ dichotomy is not sufficient to address current dilemmas where teachers need to: ‘exercise both accountability and formative conceptions of assessment’ (p314). In order to develop teacher assessment literacy, there is a need to recognise that there is not one ‘correct response’ to assessment, but a diverse range of approaches (DeLuca et al. 2016), the ongoing balance of which is dependent on purpose and context.

The Seesaw model also aims to be compatible with a ‘formative to summative’ approach to teacher assessment which will be examined in the next section: to sample the child’s performance across the whole curriculum in a manageable way that utilises information gathered formatively in the classroom, which can also be used for summative purposes.

7.4.3 Summary of validity and reliability in ‘formative to summative’ assessment

Key ideas arising from this section:

- A ‘formative to summative’ approach led to a focus on evidence in School B, which could enhance reliability, but negatively impact on validity of the formative purpose.
- School structures/criterion lists supported assessment processes but may lead to atomistic judgements and a separation of science into conceptual understanding and inquiry skills.
- Moderation was a key way schools were attempting to develop a shared understanding of assessment processes.
- The Seesaw model (Figure 7.1) was proposed which displayed a balance between validity and reliability, supported by a shared understanding of assessment and the subject, underpinned by a requirement that assessment processes are manageable.

7.5 Representation of the relationship between formative and summative assessment in the TAPS pyramid

The 'formative to summative' model proposed by Nuffield (2012) described in principle how information gathered for formative purposes (base layer) could be summarised for summative reporting purposes (all other layers above), but the model does not detail processes to enact this 'formative to summative' assessment. When the TAPS project operationalised the Nuffield model into a school self-evaluation tool (Davies et al. 2014) the processes within each layer were detailed in criteria boxes and exemplified, but the flow of information represented by the orange arrow - the process for moving up the pyramid layers, using information in the next layer - was not made explicit. In fact, where the transition from formative to summative took place was the subject of some debate and uncertainty (Davies et al. 2017). This study has been concerned with understanding the relationship between formative and summative assessment in the PSQM data and case studies, in order to inform support for teachers, including the refinement of the TAPS model and its guidance. A key outcome is to make the relationship between formative and summative assessment more explicit, and to identify where the 'formative to summative' transfer takes place in the model.

In the TAPS pyramid school self-evaluation tool (Davies et al. 2014), and subsequent iterations (Earle et al. 2015a, 2016 and 2017) the blue base layers are clearly demarcated as 'ongoing formative assessment', but the first mention of 'summative' is near the top of the pyramid in the reporting layer, hence the process of reaching a summative judgement is implied rather than explained. Based on the data in this study, three additions to the TAPS pyramid model are detailed in Figure 7.4 and then summarised in Figure 7.5:

- Noting the primarily formative use of classroom activities (Earle et al. 2017), whilst identifying that different types of activity can be more open/focused/closed (Table 7.3).
- Explicitly naming and positioning the different types of summative assessment identified in Section 7.3.2: snapshot and summary.
- Embedding of the key principles of shared understanding, validity and reliability from Figure 7.1 Seesaw to support teacher assessment literacy.

Figure 7.4 The ‘formative to summative’ process in the TAPS pyramid model (detailed)

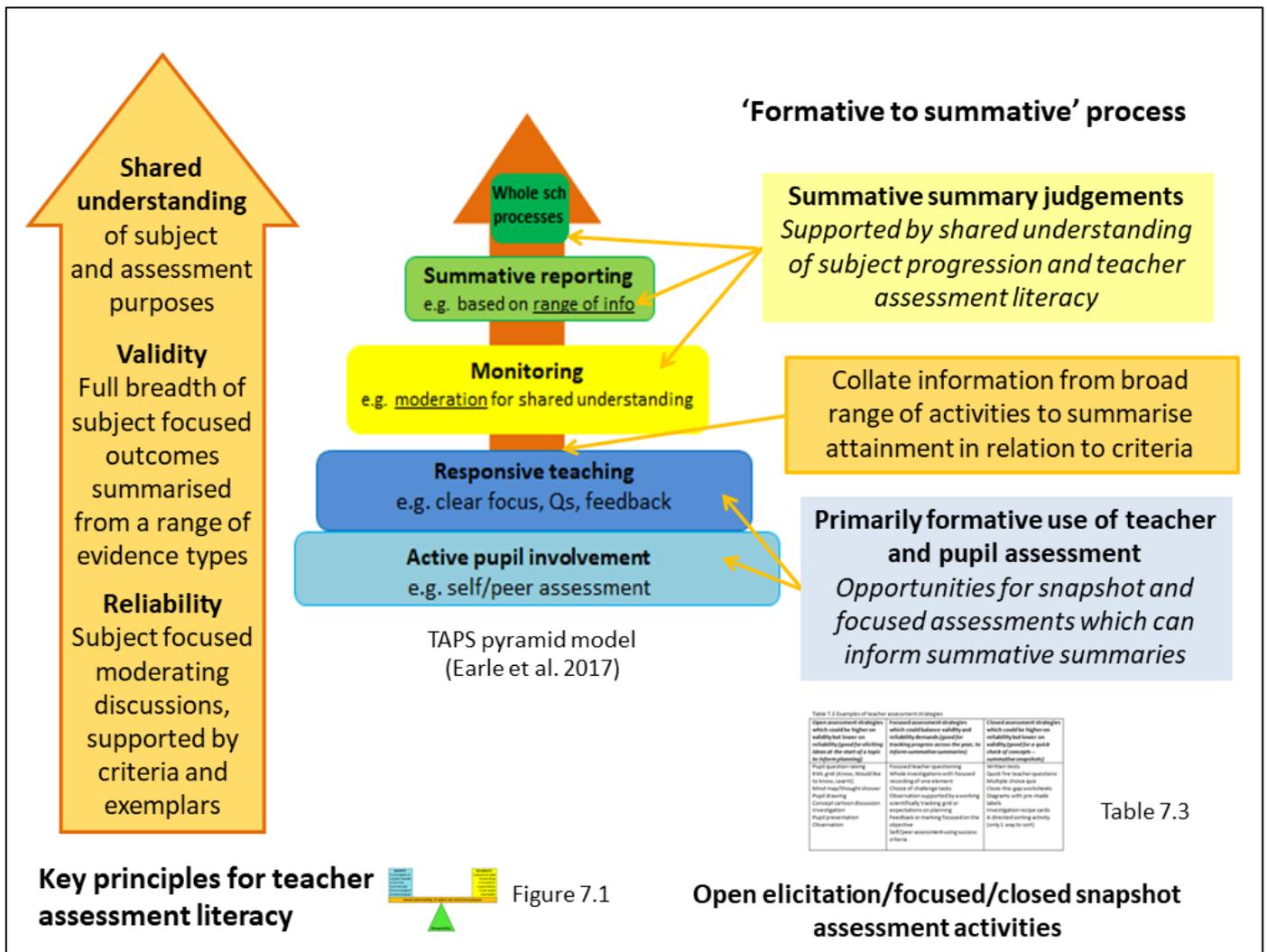


Figure 7.4 aims to draw together the thesis findings regarding teacher assessment literacy, conceptions of summative assessment and types of activity, in a way which will make the processes of ‘formative to summative’ assessment more explicit. It was noted that the TAPS pyramid model could be interpreted to mean all assessment needs to lead to summative evidence and judgements, as seen for a time in Case Study B. The ‘key principles pyramid’ (Earle et al. 2017), pictured in the centre of Figure 7.4, placed key formative messages for ‘active pupils’ and ‘responsive teachers’ within the blue base layers in an attempt to draw attention to the formative purpose of classroom activities, but it remained unclear which information should feed into the next layer. The identification of different types of activity in Table 7.3 provides guidance for practice regarding the way activities can be useful for different purposes. For example, open elicitation strategies provide information useful for

planning, whilst focused and closed snapshot activities can provide information which can later inform summaries of attainment, since they are both criterion-referenced.

In the blue layers, the formative purpose is primary, but there are opportunities for making summative judgements within individual lessons. The information gained from these 'summative snapshots' can still be used formatively, to adapt subsequent lessons, but it can also be used to inform the summary judgements aligned with the middle yellow layer. The danger of losing the formative function has not disappeared, but the emphasis on formative use has been given more prominence in the blue layers.

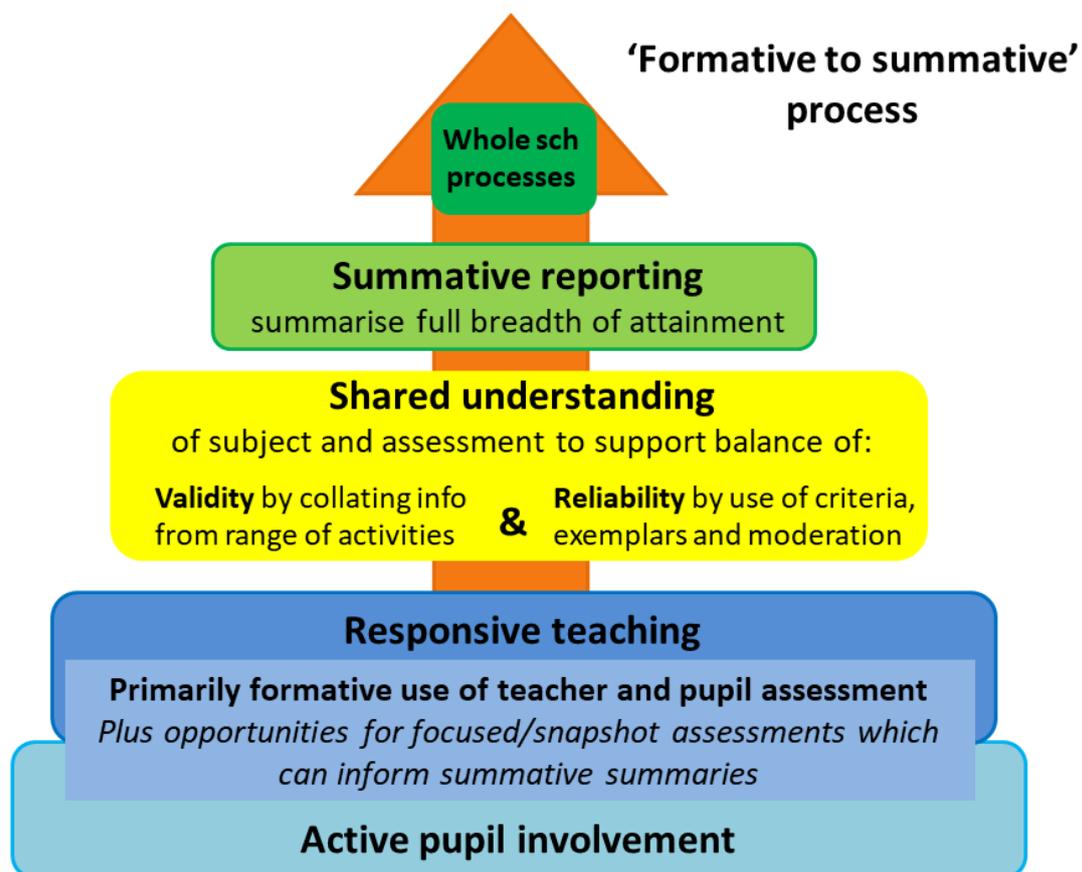
The two ways to conceptualise summative assessment, which were identified in Section 7.3.2, are explicitly positioned on the TAPS pyramid model in Figure 7.4. Opportunities for summative snapshots of attainment during classroom activities are identified within the blue pyramid layers, whilst summative summary judgements of attainment are placed in the layers above, as the result of a 'formative to summative' process. The process of information flow within the orange arrow is described as a 'collation' of assessment information, which means that summary judgements will draw on a range of information, but what this summary will look like will depend on the context, purpose or audience. For example, at the yellow 'monitoring layer' the summary may be a more detailed list of objectives which have or have not been 'met', to provide useful information for the teacher's planning for the following term. Whilst summary judgements in the green 'reporting layer' may consist of more holistic judgements for parents, which may be more 'expansive' (Lum 2015) in the way that they combine assessment information across the content domain.

The key principles from the Teacher Assessment Seesaw (Figure 7.1), which aim to support teacher assessment literacy, have been embedded within the arrow, to demonstrate that understanding of these is integral when using formative information for summative purposes. As Brown (2004) noted, simple introduction of an assessment model like the TAPS pyramid is not enough to change practice, the 'interlocked conceptions' of teachers need to be addressed in order to develop teacher assessment literacy (p314). In Figure 7.4, the key concepts of shared understanding, validity and reliability from the Seesaw are

embedded within the arrow to show that developing teacher assessment literacy in these concepts is required to be able to implement a ‘formative to summative’ process. The arrow is presented separately to the pyramid for Figure 7.4 for clarity, since the diagram contains detail regarding all of the thesis products. An alternative embedding of the key concepts can be found in Figure 7.5 below, which aims to provide more of a summary, as discussed below.

Figure 7.4 provides a detailed and annotated diagram of the process of ‘formative to summative’ assessment, drawing on all of the previously presented products. However, by drawing together all of the findings in this way, the figure has become too complex for initial presentation. Thus a simplified version is presented in Figure 7.5, which contains the same elements explored above, but draws out the key messages without linking to the other products.

Figure 7.5 The ‘formative to summative’ process in the TAPS pyramid model (summary)



The three recommendations from the bullet points above are also present in this simplified model: emphasis on primarily formative use of classroom activities; explicit naming of snapshot and summary conceptualisations of summative assessment; and the embedding of the key teacher assessment literacy principles of shared understanding and the balancing of validity and reliability. One key difference in this summary version is the expansion of the yellow pyramid layer, which has been renamed from ‘monitoring’ to ‘shared understanding’ to emphasise the key focus for practice at this layer. Validity and reliability are also explicitly mentioned, for whilst activities like moderation are contained within the detail of the TAPS pyramid boxes (Earle et al. 2015a), the reason for their inclusion was not spelled out. In order to support the development of teacher assessment literacy, it is important for teachers to, for example, identify moderation as a means of enhancing reliability in their assessments. The suggested amendments to the yellow pyramid layer effectively place the key principles from the Seesaw within that layer of the pyramid, raising its status as the key layer for the ‘formative to summative’ process. Further trialling with teachers will be necessary to consider such changes support the development of teacher assessment literacy, balancing the demands of validity and reliability in the current accountability context (DeLuca and Johnson 2017).

7.5.2 Summary of the representation of the relationship between formative and summative assessment in the TAPS pyramid

Key ideas arising from this section:

- The process and flow of information from ‘formative to summative’ was theorised but not explained in either the Nuffield (2012) or the TAPS pyramid (Davies et al. 2014, Earle et al. 2015a) models.
- Utilising the distinction between summary and snapshot summative assessment, Figure 7.4 explicitly identified the place of formative and summative assessment in the TAPS pyramid to show that summary judgements could draw upon snapshot activities.
- Teacher assessment literacy and types of activity are also linked into Figure 7.4 to signpost the further theoretical and practical products in Figure 7.1 and Table 7.3. Whilst a simplified diagram is presented in Figure 7.5 to support clarity of key messages.

7.6 Summary

After a brief summary of findings in response to the research questions, this chapter has followed three lines of discussion which have arisen from this study:

- In Section 7.3, a range of **dimensions in the conceptualisation and enactment of the relationship between formative and summative assessment** were identified and summarised in Tables 7.1 and 7.2 to provide a framework for discussion of teacher assessment. One key dimension was the conceptualisation of summative assessment as snapshot or summary. Teacher assessment strategies were categorised in Table 7.3 to provide examples for teachers of open/divergent, focused and closed/convergent assessment.
- Section 7.4 explored **validity and reliability in a 'formative to summative' approach**. A Seesaw model was presented in Figure 7.1 which presented a balance between validity and reliability, supported by a shared understanding of assessment and the subject, underpinned by a requirement that assessment processes are manageable.
- In Section 7.5 the distinction between summary and snapshot summative assessment was utilised in Figure 7.4 and 7.5 to explicitly identify **the relationship between formative and summative assessment in the TAPS pyramid model** (Davies et al. 2014, Earle et al. 2015a). The key principles of shared understanding, validity and reliability were also embedded within these diagrams to support the development of teacher assessment literacy.

From this discussion, a number of recommendations for practice have arisen, in the form of theoretical models (Tables 7.1 and 7.2 conceptualisation dimensions, Figure 7.1 Seesaw balance and Figure 7.4 and 7.5 'Formative to summative' process) and exemplification (Table 7.3 Examples of teacher assessment strategies). These will be revisited in the final chapter to make explicit the key recommendations from this study.

Chapter 8

Conclusions and recommendations

8.1 Introduction

Assessment drives the taught curriculum, defines what is valued and ‘shapes’ what is measured (Stobart 2008); it can enhance or hinder learning (Mansell et al. 2009). However, assessment is complex and has been identified as the weakest aspect of teacher practice (Black and Harrison 2010). Low assessment literacy is compounded by a lack of centralised guidance for primary teachers on how to make valid and reliable teacher assessment judgements of primary science (Turner et al. 2013). The ‘formative to summative’ pyramid model of teacher assessment, proposed by the Nuffield expert group (2012) and operationalised by TAPS (Davies et al. 2014), was designed to support teachers make principled decisions about assessment, but there was no explicit explanation of how the information gathered through formative assessment was turned into a summative judgement.

The aim of this research was to critically analyse the relationship between formative and summative assessment in action and over time, in order to develop guidance for practice to support teacher assessment in primary science. The Design-Based Research (DBR) approach resulted in new insights into the conceptualisation and enactment of the relationship between formative and summative assessment, which provided the basis for the products described in Chapter 7, including refinement of the TAPS pyramid model.

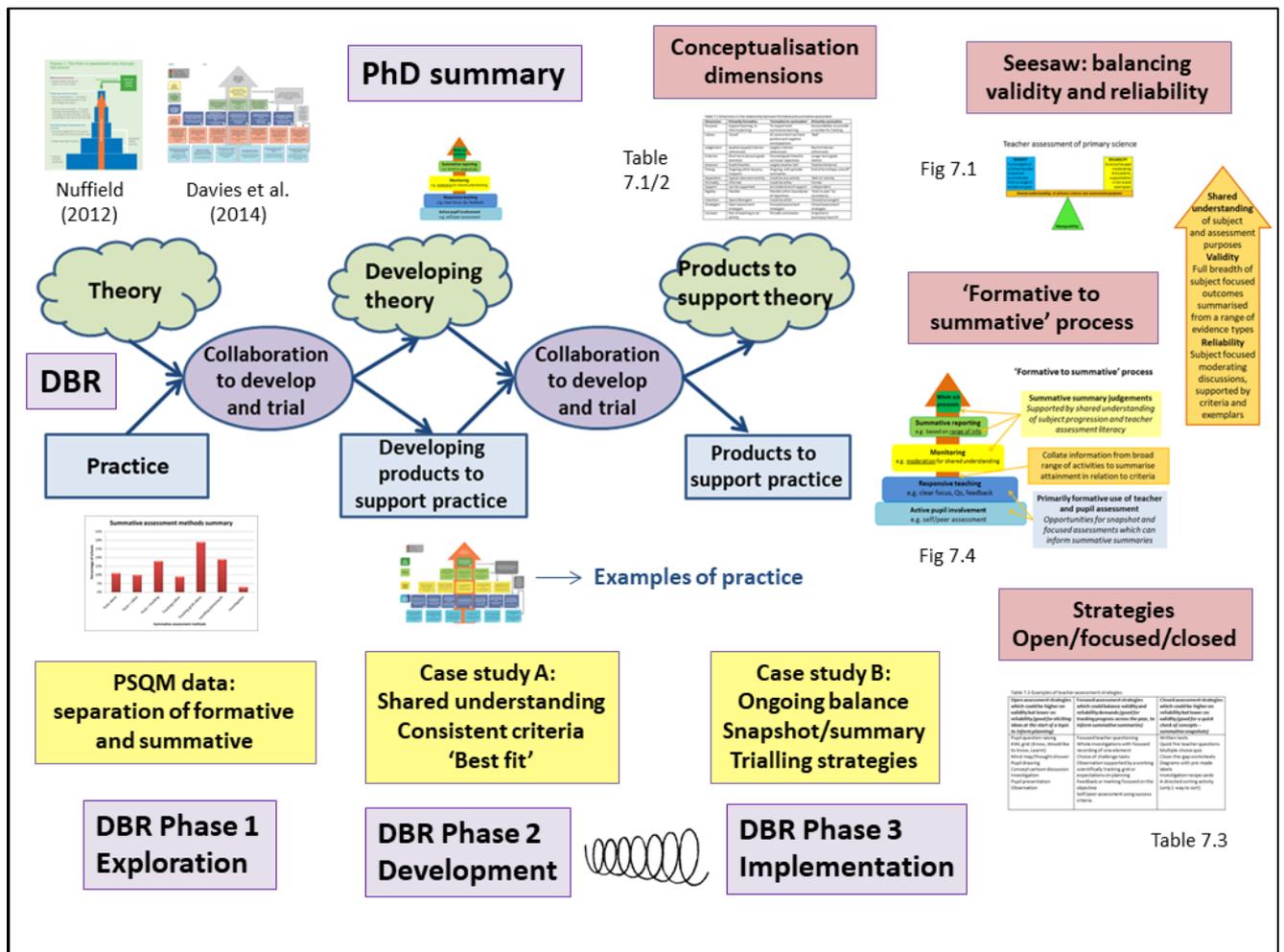
This chapter will draw together conclusions from the study and make recommendations for policy, practice and research. It will begin with a summary of key findings, followed by discussion of the advantages and limitations of the DBR process, before turning to recommendations for different audiences.

8.2 Key findings and reflections on the study

8.2.1 Key findings and products

A summary of the research is presented in Figure 8.1 to be read left to right. It contains the DBR process in the centre, the data sources in yellow boxes and the research products in pink boxes.

Figure 8.1 PhD summary



The summary presented in Figure 8.1 is designed to provide a brief overview of the research. On the left are key starting points for the research: the 'formative to summative' pyramid models of Nuffield (2012) and TAPS (Davies et al. 2014), with the latter becoming the analytical framework for the study. Within DBR Phase 1, the Primary Science Quality Mark (PSQM) database analysis found that teachers named formative and summative

assessment strategies, but these were often described separately, providing little insight into the relationship between the two. A wide range of dimensions in the conceptualisation of formative and summative assessment were identified in Case Studies A and B, which became the basis for Table 7.1, with consideration of the relationship between the two in a 'formative to summative' approach presented in Table 7.2. These conceptualisation dimensions provide a way of cataloguing and interpreting teacher understanding of assessment, with some dimensions like values and degree of separation being particularly important for a 'formative to summative' approach.

Teachers in Schools A and B attempted to enhance assessment validity by utilising information gathered for formative purposes, in line with a 'formative to summative' model, to broaden the range of types of evidence and objectives sampled. Efforts to increase the reliability of teacher assessment included moderation discussions and consistent criterion structures which supported a shared understanding of the subject. However, efforts to make teacher assessment more manageable in Case B, led at one point to closed convergent assessment based on pre-determined outcomes. Such trialling of strategies resulted in identification of the way different types of assessment activity could support different purposes, with open elicitation useful at the beginning of a topic, focused assessment supportive of ongoing tracking of inquiry skills, and closed snapshot activities supporting judgement of conceptual understanding (Table 7.3).

In both the data sources and the TAPS pyramid, the processes for making summative judgements were not transparent. A distinction between summary and snapshot summative assessment was initially considered in School A's use of a 'best fit' judgement, and explored further within Case Study B. Such a distinction was found to be useful in a 'formative to summative' model, by clarifying that summative assessment is conceived in this model as a summary, rather than a snapshot. Figure 7.4 identifies these two conceptualisations of summative assessment within the TAPS pyramid model to make the 'formative to summative' process more explicit.

Balancing the demands of validity, reliability and manageability in teacher assessment is complex and, as seen in case study B, an ongoing consideration. A Seesaw balance model of

teacher assessment was proposed to support discussion of key concepts in assessment (Figure 7.1). It was also suggested that a ‘formative to summative’ model of assessment cannot be fully implemented without a level of teacher assessment literacy: a shared understanding of both assessment and the subject is essential to be able to use formative assessment to inform summative judgements (Figure 7.5). This means that presentation of the TAPS pyramid ‘formative to summative’ model is not enough (DeLuca and Johnson 2017), an understanding of validity and reliability is necessary for the proposed flow of assessment information, which is why the key principles from the Seesaw have been placed within the orange arrow (which represents flow of information) in Figure 7.4.

A summary of these key findings and this study’s new contributions to the field is presented in Table 8.1, before moving on to reflections on the DBR process in the next section.

Table 8.1 Key findings and products

Key findings	Products	Location
Teacher conceptualisations of assessment encapsulate a wide range of dimensions related to: outcome (purpose, values, audience, intention and evidence); framework (reference, scale, assessor, timing and separation); and classroom practice (strategies, formality, rigidity and support).	Conceptualisation dimensions	Tables 7.1 and 7.2
Assessment activities can be open, focused or closed snapshots. The latter two are more criterion-referenced and can thus be useful to inform summative assessment.	Assessment strategies	Table 7.3
Teacher assessment literacy involves a shared understanding of key concepts like validity and reliability, the balance of which need to be considered over time.	Seesaw balancing validity and reliability	Figure 7.1
A ‘formative to summative’ model of assessment cannot be fully implemented without a shared understanding of key assessment concepts. Within a ‘formative to summative’ model of assessment, summative assessment is conceived as a summary judgement, which may be informed by snapshot and focused assessment activities.	‘Formative to summative’ process	Figures 7.4 and 7.5

8.2.2 Reflections on the advantages and limitations of the DBR process

During the study, DBR methodology has been a recurring theme for discussion, both in relation to the nature of DBR within a qualitative study since it emerged from experimental designs (Brown 1992), together with the practical implications of researching ‘from within’ the process. I argue that the application of DBR in this context has been fruitful, creating: *“evidence-based and ecologically valid recommendations for practice”* (McGuigan and Russell 2015: 35), but also recognise certain limitations of this study which will be explored in this section.

The case study school visits and development days were part of the TAPS project, which enabled whole school trialling of assessment strategies and multiple iterations going beyond the typical amount of time and resources available to a single researcher (Zheng 2015), with the data being analysed afresh for the PhD study. The TAPS pyramid layers were used as an analytical framework; these layers were largely stable from their inception (Davies et al. 2014), the ongoing developments being the detail within the layers (Davies et al. 2017). Using the TAPS pyramid layers as an analytical tool at the same time as trying to develop the model, could be seen to be circular, with findings seen to be both framing and confirming. However, it is an important part of the DBR process that theory and practice are developed at the same time, the **dual goals** supporting and testing each other (Design-Based Research Collective 2003). Each finding had the potential to influence the development of the theoretical model and vice versa, with the iterative cycles building the model and the practice. In addition, being part of the TAPS project supported long term and comprehensive documentation of the research in a transparent and reflexive way, with the findings regularly explained and justified to members of the team.

The **iterative cycles** of development days and school visits supported both ‘rapid prototyping’ (Bryk et al. 2010) of assessment strategies and an ongoing dialogue, both building the relationship between researcher and teacher, and acting as a driver to explore and develop practice. The DBR phases within the PhD of Exploration, Development and Implementation were used as a structural device to organise the case study data, to enable

longitudinal analysis of change over time. However, practice was continuous, where one phase ended and another began was not pre-designed, they were allocated after the process as a way of organising the data, rather than as a way of organising the action. This use of phases could be seen as part of the flexibility of DBR (Wang and Hannafin 2015), or more a naming of the 'macro'-cycles of the research rather than the 'micro' iterations (Shah et al. 2015).

Every context provides the opportunity for new insights, and so it needs to be acknowledged that this study, which considered one PSQM Round and two case studies, is based on a sample (Zheng 2015). However, it would not have been possible to present such in-depth analysis of practice over time with a broader sample, and the implications for implementation of a 'formative to summative' model may not have been visible. It is key to DBR that trialling occurs within a **real context**, 'context matters' if the research is to create products which work in practice (Barab and Squire 2004). Both theory and products to support practice must be: "*applicable and feasible in the current education system*" (Wang and Hannafin 2005: 19), which is achieved by trialling with practitioners in real contexts.

The two case study schools were similar in their context and commitment to the development of primary science, which was useful for comparative purposes, but indicates that approaches should next be trialled in a broader range of contexts; the dual goals of developing theory and products need to be ongoing. Nevertheless, the other TAPS schools, members of the research team, advisory board and other Primary Science Teaching Trust teachers, provided a valuable source of **external validation** to enable the theories presented in this study to be tested, providing a further challenge to claims of circularity. For example, when comparing School A with other schools (Earle 2015), each school had a clear structure in place (skills progression grids or 'I can' statements) which appeared to serve a similar purpose: to create a consistency or shared understanding across the school.

As well as collecting examples of practice which would be relevant for other schools, the trialling and development also impacts the context at the time (Barab and Squire 2004). One effect of the collection of examples was seen in School B, where the subject leader reported increased staff confidence that came with discussion of school strategies and

processes. However, it could be questioned whether the recognition School A was receiving, with their practice forming the basis of many of the TAPS exemplification materials (Earle et al. 2015b), could have supported complacency and a lack of development within the school, for example, with the continuation of a numerical system after the statutory levelling system had been removed. The lack of change over time in School A is impossible to reduce to the influence of one factor, but recognising the potential effect of stagnation from over-exemplification is a useful factor to be aware of for future iterations of DBR.

Being a researcher 'within' the DBR process made data collection more complex (Zheng 2015), however, it also provided a wealth of data from a range of viewpoints. Capturing continually changing practices which did not stand still to be examined was challenging, but changes in practice during implementation of a 'formative to summative' model was a key part of the study. The iterative DBR cycles did make it difficult to have time to analyse data in terms of theory before the next cycle, however, pauses in the study would have lost some of the momentum of school development and made it difficult to keep the schools engaged. The depth of data gathered in this study, through sustained collaboration and use of a range of methods to document the processes, enabled the production of two comprehensive case studies. Contextual variables, including the role of the researchers, were part of the study and an integral part of the process (Shah et al. 2015). Data-gathering at the same time as developing products with participants did lead to a blurring of the lines between researcher and practitioner, with data being collected 'in action' rather than as a separate activity, however, this can be seen as positive outcome with the teachers becoming co-researchers to enact change.

The role of the researcher distinguishes DBR from participant action research (Wang and Hannafin 2005), since it is the researcher who leads the study and provides the bridge to theory. The researcher responds to user feedback to refine theoretical models (Anderson and Shattuck 2012) and utilises the theory within developments, ensuring that it does 'real work' in real contexts (Cobb et al. 2003). The researcher is also part of the research, with **sustained collaboration** between researcher and practitioner to allow time for developments in assessment to be trialled, shared and embedded across the school. My

previous experience as a primary school teacher supported both the ongoing collaboration and relationship with schools, and the creation of products which could directly support practice. The science subject leaders in the case study schools became co-researchers but they moved from a passive role to an active role at different times (Bianchi 2017), with School A quickly taking on a leading role in providing exemplification, seeing themselves as 'experts' in the field, whilst School B did more developing and trialling first, later becoming an active partner in the presentation of strategies to others. Such professional development for participants is described by Herrington et al. (2007) as a 'societal output' of DBR, since in addition to the development of theoretical and practical products, the DBR process enhances learning for all of those involved.

8.2.3 Summary of reflections on advantages and limitations of DBR process

The following reflections have been discussed in regards to using DBR when undertaking qualitative research with teachers in this study:

- Using data collected as part of the TAPS project, together with the TAPS pyramid as an analytical framework, could lead to claims of circularity. However, there were a number of advantages to being part of the TAPS project which could help to counter such claims: long term comprehensive documentation of processes; whole school trialling of assessment approaches to gain multiple viewpoints; opportunities for external validation within and beyond the TAPS group; ongoing testing and refinement of products to support the dual DBR goals of theory and impact on practice.
- Iterative cycles of trialling in real contexts was limited by the small number of case studies, but this allowed for a depth of analysis which produced new findings for the processes involved in a 'formative to summative' approach. The impact on schools of trialling part-solutions is important to recognise for future iterations of DBR.
- The sustained collaboration between researcher and practitioners enabled long term access to the context, but the data collection 'in action' was often complicated. Nevertheless, the professional development of participants can be seen as a further output from the DBR process.

8.3 Recommendations

8.3.1 Recommendations for practice

Schools should explore their use of formative and summative assessment, considering the value they place on each. For example, if summative evidence is at the forefront, ways to raise the formative purpose should be explored. Whilst if there is little understanding of how summative summary judgements are formed, then the focus for development should be on broadening the range of evidence types by utilising moderation discussions of criteria and exemplification. The Teacher Assessment Seesaw (Figure 7.1) provides guiding principles for practice, particularly when trying to implement a ‘formative to summative’ model of assessment.

The Seesaw balance model aims to support and stimulate discussion of the purposes of assessment, but such discussions remain very abstract and removed from real practice until actual examples of teacher assessment are introduced. Subject leaders and classroom teachers involved in the TAPS project have found that the practical examples from real classrooms contained in the TAPS pyramid self-evaluation tool (Earle et al. 2015b) provide suggestions which could be immediately put to use. The Seesaw model emphasises the theory of teacher assessment, so it may need to be supplemented with examples from the TAPS pyramid to make it more accessible and immediately relevant to busy practitioners, for example, sharing strategies for recording in different ways, or exploring how to carry out moderation. However, to only look at the practical examples removed from their theoretical underpinning could lead to an adoption of strategies without an understanding of their purpose. In order to build a shared understanding of assessment, teachers need to understand their practice and judge for themselves whether changes are needed. The Seesaw model could be used to help develop assessment literacy, supporting teachers to balance validity and reliability in their teacher assessment of primary science.

The Seesaw balance model is designed to support understanding of assessment principles, but in order to consider the implications of this teachers need dedicated time for professional development and learning. In primary schools this may be a discussion which starts with the senior leadership team or in a whole school staff meeting, rather than a sole

subject leader considering changes to assessment practices alone. Moderation discussions are also a team activity; they have been highlighted as a way to improve reliability of teacher assessment, but they can also improve teaching and learning (Harlen 2007), supporting teachers to develop a better understanding of criteria and progression in a subject. Such shared understanding is key for a ‘formative to summative’ model, since assessment information gathered for primarily formative purposes is easier to utilise to inform a summative summary if it is clearly linked to subject criteria.

Summative assessment can be viewed as a snapshot or a summary, with the former providing information about pupil performance at a point in time, and the latter providing a broader view, which can be based on a range of information from classroom assessments, as in a ‘formative to summative’ model like the TAPS pyramid (Earle et al. 2015a). The process for utilising assessment information collected for primarily formative purposes is explained in Figures 7.4 and 7.5, where snapshot and focused assessments are collated to summarise attainment in relation to subject criteria.

For practical starting points for development of practice, Table 7.3 provides examples of teacher assessment strategies. They are sorted into: open strategies which elicit ideas without a predesignated answer; focused strategies which have an explicit criterion focus but allow for multiple answers; and closed strategies where one kind of answer is expected. Focused and closed strategies provide information which can be more easily summarised for summative reporting, whilst open strategies are particularly useful for children to explore their ideas and to inform planning.

Pre-determined pupil outcomes via grouping or differentiated recording may reduce validity of assessments. Focused activities, of the kind listed in the middle column of Table 7.3, need to be open enough to provide opportunities for range of outcomes, but focused on particular objectives to allow for both formative and summative uses. Teachers need to trial strategies to make them ‘work’ for their context. Change in assessment practice will take time for it is intricately entwined with teaching and learning.

Numerical systems may be useful for teacher tracking, but may signify the end of the process for pupils, thus it may not be helpful to share these with pupils. However, use of explicit criteria within planning and lessons can provide opportunities for criterion-referenced formative and summative assessment. Teachers may find that school structures, like progression charts, support explicit use of criteria and a shared understanding of development in the subject, which can support both formative and summative assessment. Nevertheless, it will be important to remember that if assessment is to enhance learning, then the formative purpose should be at the forefront and activities like high quality dialogue and self or peer assessment, which may not necessarily be utilised in summative assessments, are valuable for formative purposes.

8.3.2 Recommendations for policy

This study has demonstrated that a ‘formative to summative’ model of assessment can be implemented in primary science and the processes are clarified in Figures 7.4 and 7.5. It is a valuable approach because it can enhance validity by drawing on a wide range of classroom assessment data, balanced with enhanced reliability through a shared understanding of the subject. This takes time to develop, but the process of developing a ‘formative to summative’ approach can be fruitful in terms of staff development.

In order to develop a ‘formative to summative’ approach, teachers need access to quality exemplification materials and opportunities for moderation discussions to develop a shared understanding of criteria and standards. Teacher assessment literacy should be addressed through programmes of professional learning and Initial Teacher Education provision. Both the Teacher Assessment Seesaw (Figure 7.1) and the ‘Formative to summative’ process model (Figures 7.4 and 7.5) are proposed as tools to support such professional learning.

School accountability measures should reflect current understanding of the need to balance validity and reliability, ensuring that numerical scores and snapshot tests are not valued above assessments which draw on a wider range of data. The national inspectorate should explore with schools how judgements are made and consider the use of both formative and summative assessment.

8.3.3 Recommendations for research

Findings in this study have shed light on DBR processes when working with teachers which could provide guidance for future research. Key features and insights into DBR for researchers to consider when undertaking qualitative research with teachers were found to be: dual goals of theory and products to impact on practice; iterative cycles or phases of trialling in real contexts; and sustained collaboration between researchers and practitioners, with external validation. Also in this study, it was found that the DBR process had the potential to negatively impact school practice, for example, with collection of exemplification material perhaps contributing to stagnation of practice, and a partially-formed theoretical model could have led to a focus on evidence gathering rather than formative purpose. Thus it would be important for future DBR projects to ensure schools understand that they are taking part in a research project, where findings are hesitant, rather than implementing a fully formed programme. Nevertheless, the sustained collaboration between researchers and teachers provides fruitful opportunities to develop both theory and practice, an 'Integrated Knowledge Tradition' (Furlong and Whitty 2017) whereby research and its theories do 'real work' in education (Cobb et al. 2003). It is not acceptable to have a theory-to-practice divide, it cannot be assumed that practitioners will adopt theory through 'passive dissemination', collaboration is required to actively address educational issues together (Shah et al. 2015). This is why the products from this study have been presented as models which support both theoretical and practical understanding.

Further questions arising from this study which could form the basis of future research include:

- To what extent can the teacher conceptualisation of assessment dimensions identified in this study (Table 7.1) be applied to schools in different contexts?
- How can the sorting of assessment strategies into open/focused/closed (Table 7.3) be used to provide support for practice?
- To what extent does the distinction between summary and snapshot summative assessment support the implementation of a ‘formative to summative’ approach to assessment? (Figures 7.4 and 7.5)
- How could the Seesaw model (Figure 7.1) be used to support the development of teacher assessment literacy to balance validity and reliability in practice?
- To what extent can the TAPS pyramid model of ‘formative to summative’ assessment be applied to other contexts, subjects and phases of education?

The theoretical products from this study are: the Conceptualisation dimensions (Tables 7.1 and 7.2), the Seesaw balancing validity and reliability (Figure 7.1) and the explanation of the ‘Formative to summative’ process (Figures 7.4 and 7.5). These products will need further trialling and refinement through use in a wide range of contexts, with future iterations of DBR. The products of this study are put forward to support future research, as new contributions to the field of teacher assessment.

Appendix

References

- Abrahams, I. and Millar, R. (2008) Does practical work work? A study of the effectiveness of practical work as a teaching and learning method in school science, *International Journal of Science Education*, 30, 14: 1945-1969.
- Adelman C., Jenkins, D. and Kemmis, S. (1976) Re-thinking case study: notes from the second Cambridge Conference, *Cambridge Journal of Education*, 6:3, 139-150.
- Alexander, R. (2008) *Towards Dialogic Teaching – rethinking classroom talk*. Cambridge: Dialogos.
- Anderson, T. and Shattuck, J. (2012). Design-based research: a decade of progress in education research? *Educational Researcher*, 41(1), 16-25.
- Angrosino, M. (2012) Observation-based research. In Arthur, J., Waring, M., Coe, R. and Hedges, L. (Eds) *Research methods and methodologies in education*. London: Sage.
- ASPIRES (2013) *Aspires: Young people’s science and career aspirations, age 10 –14*. London: Kings College.
- Assessment Reform Group (1999) *Assessment for Learning: Beyond the Black Box*. Cambridge: University of Cambridge Faculty of Education.
- Barab, S. and Squire, K. (2004) Design-Based Research: Putting a Stake in the Ground, *The Journal of Learning Sciences*, 13, 1: 1-14.
- Bassey, M. (1999) *Case study research in educational settings*. Maidenhead: OUP McGraw Hill.
- Bell, D. and Ritchie, R. (1999) *Towards Effective Subject Leadership in the Primary School*. Buckingham: Open University Press.
- BERA (2011) *Ethical Guidelines*. London: BERA.
- Bianchi, L. (2017) A trajectory for the development of teacher leadership in science education. *Journal of Emergent Science*, 12: 72-83.
- Black, P. (2012) “Formative Assessment and Learning.” In Oversby, J. (Ed.) *ASE Guide to Research in Science Education*. ASE: Hatfield.
- 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.
- Black, P. and Wiliam, D. (2012) The reliability of assessments. In Gardner, J. (Ed.) *Assessment and Learning*. 2nd edition. London: Sage.
- Black, P., Harrison, C., Hodgen, J., Marshall, B. and Serret, N. (2011) Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy and Practice*, 18:4, 451-469.
- Black, P., Harrison, C., Lee, C., Marshall, B. and Wiliam, D. (2002) *Working inside the black box*. London: GL Assessment.
- Black, P., Harrison, C., Lee, C., Marshall, B., and Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham: Open University Press.
- Boaler, J. (2015) *The elephant in the classroom: helping children to learn maths*. 2nd edition. London: Souvenir Press.
- Boyle, B. and Bragg, J. (2005) No science today – the demise of primary science, *The Curriculum Journal*, 16, 4: 423-437.
- 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.
- Broadfoot, P. (2007) *An introduction to assessment*. London: Continuum.
- Brown, A. (1992) Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings, *The Journal of the Learning Sciences*, Vol. 2, No. 2, pp. 141-178.
- Brown, G. (2004) Teachers' conceptions of assessment: implications for policy and professional development, *Assessment in Education: Principles, Policy & Practice*, 11, 3, 301-318.
- Bryk, A., Gomez, L. and Grunow, A. (2010) *Getting Ideas Into Action: Building Networked Improvement Communities in Education*. Stanford: Carnegie Foundation for the Advancement of Teaching.
- Bryman, A. (2012) *Social Research Methods*, Maidenhead: OUP.
- Bryman, A. (2016) *Social Research Methods*. 5th edition. Maidenhead: OUP.
- Butler, R. (1988) Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance, *British Journal of Educational Psychology*, 58, 1-14.
- 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.
- Clarke, S. (2001) *Unlocking Formative Assessment: Practical strategies for enhancing pupils' learning in the primary classroom*. London: Hodder and Stoughton.
- Coates, D. and Wilson, H. (2003) *Challenges in Primary Science: Meeting the Needs of Able Young Scientists at Key Stage Two*. London: David Fulton Publishers.
- Cobb, P., Confrey, J, diSessa, A., Lehrer, R. and Schauble, L. (2003) The Role of Design in Educational Research, *Educational Researcher*, 32, 1, 9-13.
- 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.
- Cohen, L., Manion, L. and Morrison, K. (2011) *Research methods in education*. 7th edition. London: Routledge.
- Collins, A., Joseph, D. and Bielaczyc, K. (2004) Design Research: Theoretical and Methodological Issues, *The Journal of the Learning Sciences*, 13, 1, 15-42.
- 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.
- Connolly, S., Klenowski, V. and Wyatt-Smith, C. (2012) Moderation and consistency of teacher judgement: teachers' views, *British Educational Research Journal*, 38, 4, 593-614.
- Cowie, B. and Bell, B. (1999) A model of formative assessment in science education, *Assessment in Education: Principles, Policy and Practice*, 6, 1, 101-16.
- Cresswell, J. and Plano Clark, V. (2011) *Designing and conducting mixed methods research*. 2nd edition. London: Sage.
- Crooks, T., Kane, M. and Cohen, A. (1996) Threats to the valid use of assessments, *Assessment in Education: Principles, Policy and Practice*, 3, 3, 265-286.
- Davies, D. and McMahon, K. (2011) Smoothing the trajectory: primary-secondary transfer issues in science education. In Howe, A. and Richards, V. (Eds) *Bridging the transition from primary to secondary school*. Abingdon: Routledge.
- Davies, D., Collier, C., Earle, S., Howe, A. and McMahon, K. (2014) *Approaches to Science Assessment in English Primary Schools*. Bristol: Primary Science Teaching Trust.

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*, 22:2, 247-263.

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.

Davis, A. (1998) *The limits of Educational Assessment*. Oxford: Blackwell.

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.

DeLuca, C., LaPointe-McEwan, D. and Luhanga, U. (2016) Approaches to Classroom Assessment Inventory: A new instrument to support teacher assessment literacy, *Educational Assessment*, 21, 4, 248-266.

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

Department for Children and Families DCSF (2010) *Assessing Pupils' Progress: A teachers' handbook*. 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 (DfE) (2017) *Primary assessment in England: Government consultation*. London: DfE.

Department for Education and Employment (DfEE) (1999) *The National Curriculum Handbook for primary teachers in England*. London: DfEE.

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.

Design-Based Research Collective (2003) Design-Based Research: An Emerging Paradigm for Educational Inquiry, *Educational Researcher*, Vol. 32, No. 1, p5-8.

Digby, R. (2014) Documenting children's learning. In Davies, D., Howe, A., Collier, C., Digby, R., Earle, S. and McMahon, K. (2nd Edition) *Teaching Science and Technology in the Early Years (3-7)*. 2nd edition. London: Routledge.

- Driver, R., Guesne, E. and Tiberghien, A. (1985) *Children's ideas in science*. Buckingham: OUP.
- Dunne, M. and Maklad, R. (2015) Doing science. In Dunne, M. and Peacock, A. (Eds.) *Primary Science: A guide to teaching practice*. 2nd edition. London: Sage.
- Dunne, M. and Peacock, A. (2015) *Primary Science: A guide to teaching practice*. 2nd edition. London: Sage.
- 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. http://www.tandfonline.com/doi/full/10.1080/02635143.2014.913129#.VPgkTfmsX_E
- Earle, S. (2015) An exploration of whole school assessment systems, *Primary Science* 136: 20-22. <http://www.ase.org.uk/journals/primary-science/2015/01/136/>
- Earle, S. (2017) The challenge of balancing key principles in teacher assessment, *Journal of Emergent Science*, 12: 41-47. <https://www.ase.org.uk/journals/journal-of-emergent-science/2017/02/12/1jes-12-feb-2017-web-v3.pdf>
- Earle, S. and McMahon, K. (2017) Moderation for professional learning, *Primary Science*, 149: 28-30. <https://www.ase.org.uk/journals/primary-science/2017/09/149/4454/528-30.pdf>
- Earle, S. and Serret, N. (2012) "Children Communicating Science." In Dunne, M. and Peacock, A. (Eds.) *Primary Science: A Guide to Teaching Practice*. London: Sage.
- Earle, S., Davies, D., Collier, C., Howe, A. and McMahon, K. (2015a) *Approaches to Science Assessment in English Primary Schools: teachers' summary*. Bristol: Primary Science Teaching Trust.
- Earle, S., Davies, D., McMahon, K., Collier, C., Howe, A. and Digby, R. (2015b) *Introducing the TAPS pyramid model (interactive pdf)*. Bristol: Primary Science Teaching Trust. <https://pstt.org.uk/application/files/6314/5761/9877/taps-pyramid-final.pdf>
- Earle, S., McMahon, K., Collier, C., Howe, A. and Davies, D. (2016) The Teacher Assessment in Primary Science (TAPS) school self-evaluation tool. Bristol: Primary Science Teaching Trust. https://pstt.org.uk/application/files/8414/5761/9871/Intro_to_TAPS_sch_self_eval_pyramid_Jan_2016.pdf
- 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. (*Updated booklet, with video links*) https://pstt.org.uk/application/files/1614/8768/8133/TAPS_teacher_summary_v3_Jan_2017_to_print.pdf

- Easterday, M., Rees Lewis, D. and Gerber, E. (2014, June 23–27). Design-based research process: Problems, phases, and applications. *Learning and Becoming in Practice: The International Conference of the Learning Sciences (ICLS) 2014* (pp317-324), University of Colorado, Boulder, CO.
- Edwards, F. (2013) Quality assessment by science teachers: Five focus areas, *Science Education International*, 24, 2, 212-226.
- Evans, M. (2013) Reliability and validity in qualitative research by teacher researchers. In Wilson, E. (Ed) *School-based research: a guide for education students*. London: Sage.
- Filer, A. and Pollard, A. (2000) *The social world of pupil assessment*. London: Continuum.
- Flick, U. (2009) *An introduction to qualitative research*. 4th edition. London: Sage.
- Fullan, M. (2016) *The New Meaning of Educational Change*. 5th edition. London: Routledge
- Furlong, J. and Whitty, G. (2017) Knowledge Traditions in the Study of Education. In Whitty, G. and Furlong, J. (Eds) *Knowledge and the Study of Education: an international exploration*. Oxford: Symposium Books.
- Gardner, J., Harlen, W., Hayward, L., Stobart, G. with Montgomery, M. (2010) *Developing teacher assessment*. Maidenhead: OUP.
- Geertz, C. (1973) Thick description: towards an interpretive theory of culture. In Geertz, C. (Ed.) *The interpretation of cultures*, New York: Basic Books.
- Gibbs, G. (2012) Software and qualitative data analysis. In Arthur, J., M. Waring, R. Coe and L. Hedges (Eds.) *Research methods and methodologies in education*. London: Sage.
- Gipps, C. (1994) *Beyond Testing: Towards a theory of educational assessment*. London: Falmer Press.
- Gipps, C., Brown, M., McCallum, B. and McAlister, S. (1995) *Intuition or evidence?* Buckingham: Open University Press.
- Goldsworthy, A., Watson, R. and Wood-Robinson, V. (2000) *Developing understanding in scientific enquiry (AKSIS)*. Hatfield: Association for Science Education.
- Green, S. and Oates, T. (2009) Considering the alternatives to national assessment arrangements in England: possibilities and opportunities, *Educational Research*, 51:2, 229-245.
- Greene, J. (2010) Knowledge accumulation: three views on the nature and role of knowledge in social science. In Luttrell, W. (Ed) *Qualitative educational research*. Abingdon: Routledge.
- Guskey, T (2014) Planning professional learning, *Educational Leadership*, 71 (8), 10-16
- Guskey, T. (2002) Professional development and teacher change, *Teachers and Teaching*, 8:3, 381-391.

- Halliday, J. (2010) Educational assessment. In Bailey, R., Barrow, R., Carr, D. and McCarthy, C. (Eds.) *The Sage Handbook of Philosophy of Education*. London: Sage.
- Haney, W. and Lykes, M. B. (2010) Practice, participatory research and creative research designs. In Luttrell, W. (Ed.) *Qualitative educational research*. Abingdon: Routledge.
- Harlen, W. (1999) Purposes and procedures for assessing science process skills, *Assessment in Education*, 6(1), 129–144.
- Harlen, W. (2006) *Teaching, learning and assessing science 5-12*. 4th edition. London: Sage.
- Harlen, W. (2007) *Assessment of learning*. London: Sage.
- Harlen, W. (2008). Science as a key component of the primary curriculum: a rationale with policy implications. *Perspectives on Education: Primary Science*, 1:4–18. London: Wellcome Trust.
- 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: Global Network of Science Academies.
- Harlen, W. (2018) Learning and teaching science through inquiry. In Serret, N. and Earle, S. (Eds.) *ASE Guide to Primary Science Education*. 4th edition. Hatfield: Association for Science Education.
- Harlen, W. with Bell, D., Devés, R., Dyasi, H., Fernández de la Garza, G., Léna, P., Millar, R., Reiss, M., Rowell, P. and Yu, W. (2010) *Principles and Big Ideas of Science Education*. Hatfield: Association for Science Education.
- Harlen, W. with Bell, D., Devés, R., Dyasi, H., Fernández de la Garza, G., Léna, P., Millar, R., Reiss, M., Rowell, P. and Yu, W. (2015) *Working with big ideas of science education*. Trieste, Italy: Science Education Programme.
- Harlen, W. and Qualter, A. (2014) *The Teaching of Science in Primary Schools*. Abingdon: Routledge.
- Harrison, C. and Howard, S. (2009) *Inside the Primary Black Box*. London: GL Assessment.
- Harrison, C. and Howard, S. (2010) Issues in primary assessment: 1 Assessment purposes, *Primary Science*, 115, 5-7.
- Hartas, D. (Ed) (2010) *Educational Research and Inquiry*. London: Continuum.
- Hattie, J. (2009) *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. Abingdon: Routledge.
- Hedges, L. (2012) Design of empirical research. In Arthur, J., Waring, M., Coe, R. and Hedges, L. (Eds) *Research methods and methodologies in education*. London: Sage.

Herrington, J. and Reeves, T.C. (2011). Using design principles to improve pedagogical practice and promote student engagement. In Williams, G., Statham, P., Brown, N. and Cleland, B. (Eds.) *Changing Demands, Changing Directions*. Proceedings ascilite Hobart 2011. (pp.594-601).

Herrington, J., McKenney, S., Reeves, T. and Oliver, R. (2007). Design-based research and doctoral students: Guidelines for preparing a dissertation proposal. In Montgomerie, C. and Seale, J. (Eds.) *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007* (pp. 4089-4097). Chesapeake, VA: AACE.

Hitchcock, G., & Hughes, D. (1995). *Research and the teacher: A qualitative introduction to school-based research*. 2nd edition. London: Routledge.

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

Homan, R. (2002) The principle of assumed consent: the ethics of gatekeeping. In McNamee, M. and Bridges, D. (Eds.) *The ethics of educational assessment*. Oxford: Blackwell.

Howe, A., Davies, D., McMahon, K., Towler, L., Collier, C. and Scott, T. (2009) *Science 5-11: a guide for teachers*. 2nd edition. Abingdon: David Fulton Publishers.

Isaacs, T., Zara, C. and Herbert, G. with Coombs, S. and Smith, C. (2013) *Key concepts in Educational Assessment*. London: Sage.

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. (2012) *Assessing Learning in the Primary Classroom*. Abingdon: Routledge.

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

Kelly, A. (2003) Research as design, *Educational Researcher*, Vol. 32, No. 1, p3-4.

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

Klenowski, V. and Wyatt-Smith, C. (2014) *Assessment for Education*. London: Sage.

Knight, S., Buckingham Shum, S. and Littleton, K. (2014) Epistemology, assessment, pedagogy: where learning meets analytics in the middle space, *Journal of Learning Analytics*, 1(2) pp. 23–47.

Kvale, S. and Brinkmann, S. (2009) *InterViews: Learning the craft of qualitative research interviewing*. 2nd edition. Los Angeles: Sage.

Lincoln, Y. and Guba, E. (1985) *Naturalistic inquiry*. California: Sage.

- Loughland, T. and Kilpatrick, L. (2013) Formative assessment in primary science, *Education 3-13: International Journal of Primary, Elementary and Early Years Education*, 43, 128-141.
- Lum, G. (2015) Introduction, Afterword. In Davis, A. & Winch C. with Lum, G. (Eds.) (2015) *Educational Assessment on Trial*. London: Bloomsbury.
- Luttrell, W. (2010) *Qualitative educational research*. Abingdon: Routledge.
- Mansell, W., James, M. and the Assessment Reform Group (2009) *Assessment in schools: fit for purpose?* London: Teaching and Learning Research Programme.
- Marking Policy Review Group (2016) *Eliminating unnecessary workload around marking: Report of the Independent Teacher Workload Review Group*. London: DfE.
- Mason, J. (1996) *Qualitative researching*. London: Sage.
- Mawby, T. and Dunne, M. (2012) Planning for Assessment for Learning. In Dunne, M. and Peacock, A. (Eds.) (2012) *Primary Science: A Guide to Teaching Practice*. London: Sage.
- Maxwell, J. (2010) Validity: how might you be wrong? In Luttrell, W. (Ed.) *Qualitative educational research*. Abingdon: Routledge.
- McGuigan, L. and Russell, T. (2015) Using multimodal strategies to challenge early years children's essentialist beliefs, *Journal of Emergent Science*, 9, p35-41.
- McMahon, K. and Davies, D. (2003) Assessment for inquiry: supporting teaching and learning in primary science, *Science Education International*, 14, 4: 29-39.
- Mears, C. (2012) In-depth interviews. In Arthur, J., Waring, M., Coe, R. and Hedges, L. (Eds.) *Research methods and methodologies in education*. London: Sage.
- Messick, S. (1989) Meaning and values in test validation: the science and ethics of assessment, *American Educational Research Association*, 18, 2, 5-11.
- Millar R. (2010) Practical work. In J. Osborne and J. Dillon (Eds.) *Good practice in science teaching: what research has to say*. Maidenhead: Open University Press.
- Mishler, E. (2010) Validation in inquiry-guided research. In Luttrell, W. (Ed) *Qualitative educational research*. Abingdon: Routledge.
- Mortimer, E. and Scott, P. (2003) *Meaning making in secondary science classrooms*. Maidenhead: Open University Press.
- Murphy, C., Lundy, L., Emerson, L. and Kerr, K. (2013) Children's perceptions of primary science assessment in England and Wales, *British Educational Research Journal*, 39, 3, 585-606.
- Murphy, P. (Ed) (1999) *Learners, Learning and Assessment*. London: Sage.

- Naylor, S. and Keogh, B. (2000) *Concept Cartoons in Science Education*. Sandbach: Millgate House Publishers.
- Naylor, S., Keogh, B. and Goldsworthy, A. (2005) *Active Assessment: thinking, learning and assessment in science*. Sandbach: Millgate House Publishers.
- Newby, P. (2010) *Research Methods for Education*, Harlow: Pearson.
- Newton, P. (2009). 'The reliability of results from national curriculum testing in England', *Educational Research*, 51, 2, 181–212.
- Noyes, A. (2004) Learning landscapes, *British Educational Research Journal*, 30 (1): 27-41.
- Nuffield Foundation (2012) *Developing policy, principles and practice in primary school science assessment*. London: Nuffield Foundation.
- Ofsted (2012a) School A inspection report - reference withheld to preserve anonymity.
- Ofsted (2012b) School B inspection report - reference withheld to preserve anonymity.
- Ofsted (2013) *Maintaining Curiosity: A survey into science education in schools*. Report No. 130135. Manchester: Ofsted.
- Ollerenshaw, C. and Ritchie, R. (1993) *Primary science: making it work*. London: David Fulton Publishers.
- Ollerenshaw, C. and Ritchie, R. (1997) *Primary Science: Making it Work*. 2nd edition. London: David Fulton.
- Oppenheim, A. (1992) *Questionnaire design, interviewing and attitude measurement*. 2nd edition. London: Pinter Publishers.
- Piaget, J. (1961) A genetic approach to the psychology of thought, *Journal of Educational Psychology*, 52, 151-161.
- Pollard, A. (2014) *Reflective teaching in schools*. 4th edition. London: Bloomsbury.
- Porritt, V. (2014) Evaluating the impact of professional learning. In Crowley, S. (Ed) *Challenging Professional Learning*. Abingdon: Routledge.
- Roberts, R. and Gott, R. (2006) Assessment of performance in practical science and pupil attributes, *Assessment in Education: Principles, Policy & Practice*, 13:01, 45-67.
- Robson, C. (2011) *Real world research*. 3rd edition. Chichester: Wiley.
- Rowe, M. (1972) Wait time and rewards as instructional variables: their influence on language, logic and fate control. Paper presented at *National Association for Research in Science Teaching*, Columbia University, New York.

- Russell, T., and Harlen, W. (1990) *Assessing Science in the Primary Classroom: Practical Tasks*. London: Paul Chapman Publishing.
- Russell, T., Bell, D., Longden, K. and McGuigan, L. (1993) *Primary SPACE Research report: Rocks, soil and weather*. Liverpool: Liverpool University Press.
- Sadler, R. (1989) Formative assessment and the design of instructional systems, *Instructional Science*, 18, 119-144.
- Shah, J., Ensminger, D. and Their, K. (2015) The Time for Design-Based Research is Right and Right Now, *Mid-Western Educational Researcher*, 27, 2:152-171.
- Sharpe, R. (2004) How do professional learn and develop? Implications for staff and educational developers. In Baume, D. and Kahn, P. (Eds) *Enhancing staff and educational development*. London: RoutledgeFalmer.
- Shavelson, R., Phillips, D., Towne, L. and Feuer, M. (2003) On the Science of Education Design Studies, *Educational Researcher*, 32, 1, 25-28.
- Silverman, D. (2011) *Interpreting qualitative data: a guide to principles of qualitative research*. 4th edition. London: Sage.
- Simons, H. (1996) The paradox of case study, *Cambridge Journal of Education*, Vol. 26, Issue 2, 225-240.
- Simons, H. (2009) *Case study research in practice*. London: Sage.
- Simons, H. and Usher, R. (2000) *Situated Ethics in Educational Research*. London: RoutledgeFalmer.
- Sizmur, S. and Sainsbury, M. (1997) Criterion referencing and the meaning of national curriculum assessment, *British Journal of Educational Studies*, 45, 2, 123–140.
- Sprague, J. (2010) Seeing through Science: Epistemologies. In Luttrell, W. (Ed.) *Qualitative educational research*. Abingdon: Routledge.
- Stake, R. (2006) *Multiple Case Study Analysis*. New York: Guilford Press.
- Standards and Testing Agency (2015) *Interim teacher assessment frameworks at the end of key stage 2*. London: STA.
- Standards and Testing Agency (2016) *2016 teacher assessment exemplification: end of key stage 2 Science*. London: STA.
- Standards and Testing Agency (2017) *Teacher assessment frameworks at the end of key stage 2: For use in the 2017 to 2018 academic year*. London: STA.
- Standish, P. (2007) Rival conceptions of the philosophy of education, *Ethics and Education*, 2, 2, 159-171.

- Stobart, G. (2008) *Testing Times: The Uses and Abuses of Assessment*. London: Routledge.
- Stobart, G. (2009) Determining validity in national curriculum assessments, *Educational Research*, 51, 2, 161–179.
- Stobart, G. (2012) Validity in formative assessment. In Gardner, J. (Ed.) *Assessment and Learning*, 2nd Edition, London: Sage.
- Stoll, L. Bolam, R., McMahon, A., Wallace, M and Thomas, S. (2006) Professional learning communities: a review of the literature, *Journal of Educational Change*, 7: 221-258.
- Swaffield, S. (2011) Getting to the heart of authentic Assessment for Learning, *Assessment in Education: Principles, Policy & Practice*, 18:4, 433-449.
- Taras, M. (2005) Assessment – summative and formative – some theoretical reflections, *British Journal of Educational Studies*, 53, 4, 466-478.
- Taras, M. (2007) Machinations of assessment: metaphors, myths and realities, *Pedagogy, Culture and Society*, 15:1, 55-69.
- Teddlie, C. and Yu, F. (2007) Mixed Methods Sampling: A Typology With Examples, *Journal of Mixed Methods Research*, 1, 77-100.
- Tharp, R. and Gallimore, R. (1988) *Rousing Minds to Life: Teaching, Learning and Schooling in Social Context*. New York: Cambridge University Press.
- Torrance, H. and Prior, J. (1998) *Investigating formative assessment*. Maidenhead: OUP.
- 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.
- Ulichny, P. and Schoener, W. (2010) Teacher-researcher collaboration from two perspectives. In Luttrell, W. (Ed) *Qualitative educational research*. Abingdon: Routledge.
- Usher, R. (2000) Deconstructive happening, ethical moment. In Simons, H. and Usher, R. (Eds.) *Situated Ethics in Educational Research*. London: RoutledgeFalmer.
- Vygotsky, L. (1978) *Mind in Society*. Cambridge, MA: Harvard University Press.
- Walker, R. (1983) Three Good Reasons for not Doing Case Studies in Curriculum Research, *Journal of Curriculum Studies*, 15:2, 155-165.
- Wang, F. and Hannafin, M. (2005) Design-Based Research and Technology-Enhanced Learning Environments, *Educational Technology Research and Development*, 53(4), 5–23.
- 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.
- Wellcome Trust (2014) *Primary Science: Is It Missing Out? Recommendations for reviving primary science*. London: Wellcome Trust.
- Wenham, M. and Ovens, P. (2010) *Understanding primary science*. 3rd edition. London: Sage.
- 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.
- William, D. (2003). 'National curriculum assessment: how to make it better', *Research Papers in Education*, 18, 2, 129–136.
- William, D. (2011) *Embedded formative assessment*. Bloomington: Solution Tree Press.
- William, D. and Black, P. (1996) Meaning and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, 22, 5, 537-549.
- Wilshaw, M. (2016) *HMCI's commentary: science and foreign languages in primary school*. Available at: <https://www.gov.uk/government/speeches/hmcis-monthly-commentary-may-2016>
- Yin, R (2014) *Case study research: design and methods*. 5th edition. London: Sage.
- Zheng, L. (2015) A systematic literature review of design-based research from 2004 to 2013, *Journal of Computers in Education*, 2(4): 399–420.

Chapter 3 Appendices

Appendix 3A: Collection of Assessment Samples

We have previously asked the attendees at Cluster day 1 to provide us with examples of their assessment approach in practice. Below is a checklist to help structure and track this process.

Document	Exists?	Collected?
Policy (e.g. assessment section of science policy, science section of assessment policy)		✓
Observations (e.g. teacher observation notes, annotated photos etc.)		✓
Annotated samples of work in science from Y1 to Y6		✓
Assessment tools (e.g. Concept cartoons/concept maps/floorbooks/KWL grids etc.)	✓	
Examples of any test used (e.g. end of unit tests, SATS, Rising Stars etc.)	✓	
Examples of any pupil self-assessment (e.g. pupil comments on work)		✓
Annotated planning (e.g. indicating which pupils reached objectives)		✓
Tracking grid (e.g. APL, traffic light systems, computerised systems)	✓	
Examples of target-setting for groups or individuals in science		✓
Examples of reporting to parents on science	✓	
Other (please specify):		

It will be useful to have these examples to hand when interviewing the science subject leader to illustrate answers with examples from school practice.

Appendix 3B: TAPS cluster day 1 formative/summative written task Oct13

Name:

School:

Role:

Years as a teacher:

What does 'formative' assessment mean to you?

What does 'summative' assessment mean to you?

Appendix 3C: Impact Questionnaire to TAPS project teachers June 2015

1. What is/are your role(s) in school? (e.g. Y4 classteacher and science subject leader)
2. What has your involvement in the TAPS project been? (e.g. From the beginning of the project as science subject leader./This is the second meeting I have attended.)
3. What have you personally gained being involved with the TAPS project?
4. In what ways has your school changed the ways you assess children's progress in science as a result of the TAPS project?
5. Have you or your school made use of the TAPS pyramid tool? Yes/No
If yes – please explain how you have used it. If no – Please explain why not.
6. Please rate the overall usefulness of the TAPS pyramid tool from 1 (not at all useful) to -5 (extremely useful).

1 2 3 4 5
7. Have you shared ideas or resources from the TAPS Project with your colleagues in school? Yes/No. If yes, which aspects did you share and what was their response?
8. How have you shared TAPS with your colleagues beyond your own school? Yes/No. If yes, which aspects did you share and what was their response
9. What would do you see as the next steps for the project?
10. What would you like as the focus for tutor visits to your school?
11. Please make any further comments here.

Appendix 3D: Lesson observation schedule/framework, School visit 1 Nov13

Observations of science lesson(s)

Observe excerpts from lessons (no need to observe the whole) and identify:

- Examples of teachers making/recording assessment judgements or gathering evidence
- Opportunities for assessment (whether or not taken)

The focus of the observations is formative use of assessment in the classroom, but evidence gathered for formative purposes can also be noted.

If possible, collect copy of lesson plan and any associated pupil work.

Opportunities for assessment identified in lesson plan:

Proforma for lesson observation (categories drawn from Harlen (2013))

Year group(s):	Time from: to:	Lesson focus:
-----------------------	-----------------------	----------------------

	Assessment opportunities taken	Further assessment opportunities
Teachers involve students in discussing learning goals and the standards to be expected in their work		
Teachers gather evidence of their students' learning through questioning/ discussion		
Teachers gather evidence of their students' learning through observation		
Teachers gather evidence of their students' learning through study of products relevant to the learning goals		
Teachers use assessment to advance students' learning by adapting the pace, challenge and content of activities		
Teachers use assessment to advance students' learning by giving feedback to students about how to improve		
Teachers use assessment to advance students' learning by providing time for students to reflect on and assess their own work		
Other (please specify):		

If possible, have a brief chat with teacher after the lesson:

Share completed lesson observation proforma and ask for their comments/validation of observations made.

Teachers' comments on observation and assessment opportunities taken/missed:
--

Ask the teacher how the approach taken during the lesson fits in with their overall approach to assessment in science:

Teachers' comments on the school's approach and how it is enacted within the classroom:

3E: Lesson observation schedule/framework, School visit 3, March 2014

Proforma for lesson observation (categories drawn from Harlen (2013) and Short (2014))

Year group(s):	Time from: to:	Lesson focus:	
Teachers' role	Observed evidence	Pupils' role	Observed evidence
Teachers involve students in discussing learning goals and the standards to be expected in their work		Pupils Identify with adults their learning needs	
Teachers gather evidence of their students' learning through questioning/discussion		Pupils focus on key aspects of the tasks with reference to success criteria	
Teachers gather evidence of their students' learning through observation		Pupils articulate their difficulties	
Teachers gather evidence of their students' learning through study of products relevant to the learning goals		Pupils are expecting/demanding feedback on their efforts	
Teachers use assessment to advance students' learning by adapting the pace, challenge and content of activities		Pupils collaboratively identify next steps in learning	
Teachers use assessment to advance students' learning by giving feedback to students about how to improve		Pupils evaluate their own and others' work against known criteria	
Teachers use assessment to advance students' learning by providing time for students to reflect on and assess their own work		Pupils make improvements in response to suggestions given	
Other (please specify):			

3F: Interview with science subject leader, school visit 1, November 2013

Semi-structured (main questions underlined with sub-questions as prompts). Make notes and voice-record if possible, referring to samples as appropriate:

General questions

- What do you see as the purposes of assessment in science?
- Do colleagues have a shared understanding of assessment – have you discussed assessment in science?
- How do colleagues develop their understanding and practice of assessment?
- How do you know about assessment in school – how are you supported in finding out?
- How does the ethos of the school affect your approach to assessment?

Using the 'Flow of Assessment Data' model from Nuffield report (2012) and referring to the examples collected, work upwards through the model levels with science co-ordinator.

Level 1 How do colleagues feedback to children about their progress and attainment in science? What informs their feedback?

- How do colleagues make and record observations of children?
- Who is involved in making observations?.
- Do approaches vary across age ranges?
- Who is involved in making judgements about learning and next steps?
- What opportunities are there for self/peer assessment in science?

Level 1/2 How do colleagues monitor children's progress? What information do they draw on?

- What is the period that assessments are made, each week? Each unit?
- What sort of range of activities/tasks do teachers use to gather assessment evidence?
- Do you assess children as individuals, groups or whole classes?
- In what ways are skills and knowledge (conceptual understanding) assessed? Are they assessed separately?
- Do you moderate summative assessments?

How do colleagues record evidence of progress in order to report upon it?

- Do you use samples of children or 'marker children'?
- How do you ensure assessments are reliable and consistent?
- Are children aware of the criteria used to judge their learning in science and how judgements are made?
- Are children aware of the range of evidence used to judge their learning in science?
- How are judgements of children's progress in science fed back to them (e.g. through target-setting)?

Level 2 How do colleagues report on children's progress in science to parents and other colleagues?

Level 3 How do colleagues contribute to school performance data for science?

- Do you conduct summative assessment tasks across classes?
- Do you use end of key stage testing?
- How do you transfer information from one system to another?
- Are the data transferable to other schools?

3G: Interview questions for science subject leader, June 2016

1. What has been your role in the TAPS project?
2. What have been the main benefits for your school of being involved in the TAPS project?
3. What changes to science assessment have you made across the school since the beginning of the project?
4. Please outline the relationship between formative and summative assessment of science in your school
5. Can you give any examples of how science assessment in your school has become more valid?
6. Can you give any examples of how science assessment in your school has become more reliable?
7. Can you give any examples of how science assessment in your school has become more manageable?
8. Can you give any examples of how science assessment in your school has made a positive impact on pupils' learning?

Appendix 3H: TAPS project consent forms

Extracts from school agreement and teachers' consent to participate forms:

School Agreement (completed by Head teacher)

As a school we undertake to enable the above teacher to participate in TAPS project cluster days (supply cover provided) and to support them in carrying out science assessment development work in their classrooms.

We will nominate a person (Head or science subject leader) to check and provide permission for any material from our school which is to be shared on the Primary Science Teaching Trust website (school name only). If there are photos or videos of children then we would undertake to gain permission from parents and children using the appropriate TAPS permission forms.

Teacher consent to participate

As part of our commitment to ethical practice in research, we would like to request your informed consent to participate in the TAPS project (2013-16). The aim of the project is to develop support for teacher assessment in primary science which is valid, reliable, manageable and has a positive impact on children's learning.

During the project, we may request:

- *Interviews with you to reflect upon your school's approach to science assessment.*
- *Samples of science assessments, for example, children's work or teacher records.*
- *Permission to observe lessons to explore science assessment in action or to pilot new approaches with you.*

Although, owing to the nature of some of the above data, we cannot guarantee that it will be stored anonymously, we do undertake to preserve your anonymity in any reports to our funders, publications or other material emerging from the project. The data will be stored securely and only shared within your school and the research team. It will not be passed onto any third parties without your permission. You have the right to withdraw from the research at any time, in which case data relating to your involvement will be destroyed.

Please sign and date below to indicate your willingness to participate.

Chapter 4 Appendices

Appendix 4A: PSQM Round 4 data

Link to Google folder of anonymised SL submissions:

<https://drive.google.com/drive/folders/0B7CbtsNKIEZcNHJyLURTY2VHdk0?usp=sharing>

102 applications for Round 4 PSQM:

Included (anonymous school codes in black type) - **91 school submissions were used in the analysis.**

Removed (red type) - 10 schools who had not completed the submission (they had not written their C2 reflections) and 1 school which was not based in England.

Hub	School code Round.hub.school							
1	R4.1.1	R4.1.2	R4.1.3	R4.1.5	R4.1.6	R4.1.7	R4.1.8	
	R4.1.4	Deferred to R5						
2	R4.2.1	R4.2.2	R4.2.4	R4.2.5	R4.2.6	R4.2.8		
	R4.2.3	Withdrawn		R4.2.7	Deferred to R5			
3	R4.3.1	R4.3.2	R4.3.4					
	R4.3.3	INCOMPLETE						
4	R4.4.1							
	R4.5.1	R4.5.2	R4.5.3	R4.5.4	R4.5.5			
6	R4.6.1	R4.6.2						
7	R4.7.1	Discounted because not in England						
8	R4.8.2							
	R4.8.1	INCOMPLETE						
9	R4.9.1	R4.9.2	R4.9.3	R4.9.4	R4.9.5			
10	R4.10.1	R4.10.2	R4.10.3	R4.10.4	R4.10.5	R4.10.6	R4.10.7	R4.10.8
11	R4.11.1	R4.11.2	R4.11.3					
12	R4.12.1	INCOMPLETE						
13	R4.13.1							
14	R4.14.1	R4.14.2	R4.14.3	R4.14.4	R4.14.5	R4.14.6	R4.14.7	R4.14.8
15	R4.15.1	R4.15.2	R4.15.3	R4.15.4				
16	R4.16.3							
	R4.16.1	INCOMPLETE		R4.16.2	INCOMPLETE			
17	R4.17.1	R4.17.2	R4.17.3					
18	R4.18.1		R4.18.3	R4.18.4	R4.18.5	R4.18.6	R4.18.7	
	R4.18.2	INCOMPLETE						
19	R4.19.1	R4.19.2	R4.19.3	R4.19.4	R4.19.5	R4.19.6	R4.19.7	R4.19.8
20	R4.20.1	R4.20.2	R4.20.3	R4.20.4				
21	R4.21.1	R4.21.2		R4.21.4	R4.21.5	R4.21.6		
	R4.21.3	INCOMPLETE						
22	R4.22.1	R4.22.2	R4.22.3					
23	R4.23.1	R4.23.2	R4.23.3	R4.23.4				
24	R4.24.1	R4.24.2	R4.24.3					

Appendix 4B: PSQM initial coding

After an initial reading of the first 30 PSQM submissions, it became clear that the data could be coded for formative and summative assessment, in answer to RQ1. Atlas.TI software was used to support this:

- All 91 documents were imported into the software.
- Initial codes were created from the RQs:

Theory-led codes: *formative, summative, purpose.*

- Whilst reading each submission, sentences pertaining to the codes were highlighted.
- New codes were added as they appeared in the data:

Emergent codes: *APP (assessing pupil progress), tracking, tests, summative named by teacher, summative other, moderation, next step identified as moderation, next step for testing, AfL /formative named by teacher, elicitation strategies, learning objective, marking, gaps in learning, self-assessment, peer-assessment*

The codes were initially grouped into summative and formative, with notes kept in Excel regarding use of the code/thoughts arising, as shown in the tables below:

Initial coding of **summative assessment** in the 91 submissions:

Emergent codes for summative	Frequency of comments*	Notes
APP (assessing pupil progress)	75	Higher than expected
Tracking	58	Many different applications for tracking
Tests	45	Many say tests, but often in combination with other strategies
Summative named by teacher	30	Many teachers use the word
Summative other	35	Inc end of unit/yr, passing onto next T
Moderation	29	Lots trialling this
Next step identified as moderation	15	Lots plan to do this next
Next step for testing	11	Some want to change their use of tests

*Frequency of comments, could be more than one comment within a school's submission, so this is not the number of schools.

The initial summative codes did not seem to represent the way schools were combining summative approaches, so the data was re-examined in a **second level of analysis** to classify different combinations and ensure that each school's practice was included (see Appendix 4C).

Initial coding of **formative assessment** in the 91 submissions:

Code	Frequency	Notes
AfL /formative named by teacher	48	AfL/formative used interchangeably
Elicitation strategies	69	Wide range – need to sub-code these
ID gaps in learning/move learning forward	32	Shows formative purpose
Learning obj/success criteria	25	Making expectations explicit
Marking	30	Not all mention – do not see as 'assessment'?
Peer assessment	24	Less than self-assessment
Self-assessment	45	How using?

Some of the codes were too broad, so the data was re-examined in a **second level of analysis** to sub-code the elicitation strategies (Appendix 4D).

Appendix 4C: PSQM second level of analysis for summative assessment

The initial codes (Appendix 4B) did not seem to represent the way schools were combining summative approaches, so the data was re-examined in a **second level of analysis** to classify different combinations and ensure that each school's practice was included. This was done by re-reading the summative coding for each school and classifying it using the emergent 'detailed categories' below:

Detailed categories (listed in Excel)	Number of schools	Notes
Tests	10	described 'tests'
Tests to back up T judgement	3	'to back up', support TA (not say how)
Combined tests + other	6	'Other'=Use of I can statements, end of unit grid, written/verbal assessment
Investigation/focused AT1	3	Mentioned focused SC1/enquiry tasks
Levelling work (at end of unit)	13	Mentioned 'levelling work' (not say how)
Combined tracking grid/APP + tests	16	Mentioned tracking grids and tests
Combined tracking grid/APP + other	8	Mentioned tracking grid plus 'other' – above
APP Tracking grid	11	Mentioned APP
Other Tracking grid	15	Mentioned tracking grid, other than APP
Levels on planning	4	Mentioned 'levels on planning'
No mention of sum	2	-
Total	91	

(Presented in graph form in Figure 4.1)

Themes arising: (Presented in graph form in Figure 4.2)

Summary categories	No. of sch	% of sch	Notes
Tests alone	10	11%	Only mentioned tests
Tests + other	9	10%	Inc back up judgemt and I can etc
Tests + tracking	16	18%	Combined tests and tracking
Tracking+other	8	9%	Tracking plus end of unit grid etc
Tracking grids alone	26	29%	Only mentioned tracking
Levelling-plans/work	17	19%	TA Level on plans or level work at end
Investigation	3	3%	Foc enq task/investigation
Combination of methods	33	37%	More than one summative method

% out of 89 because 2 did not specify

Appendix 4D: PSQM second level of analysis for formative assessment

Atlas.TI was used for **initial coding** of formative assessment in the 91 submissions:

Code	Frequency	Notes
AfL /formative named by teacher	48	AfL/formative used interchangeably
Elicitation strategies	69	Wide range – need to sub-code these
ID gaps in learning/move learning forward	32	Shows formative purpose
Learning obj/success criteria	25	Making expectations explicit
Marking	30	Not all mention – do not see as 'assessment'?
Peer assessment	24	Less than self-assessment
Self-assessment	45	How using?

Some of the codes were too broad, so the data was re-examined in a **second level of analysis** to sub-code the elicitation strategies. These sub-codes were then grouped into themes:

Themes for formative assessment

Elicitation strategies PSQM Round 4		
Paper/task based	Tests/asst shts	11
	Quiz/game	8
	Mind/concept map	9
	KWL grids	7
	Own Qs	5
	Concept cartoon/'active asst'	12
	Self assessment at beg	2
Obs	Postit, photo, vid	16
	Observation of task	19
Collab	Floorbook	2
	Investn/practical task	5
	Grp challenge	5
	Presn/ppt	6
	Drama/role play	3
Teacher led talk	Talk partner/pair-share	7
	Discussion	8
	Questioning	29
Other elicitation		7

Chapter 5 Appendices

Appendix 5A: School A case record

The data for School A was collected June 2013 - June 2015 and included 6 TAPS cluster days, 5 school visits and one PSQM application.

Link to anonymised case record folder containing each item:

<https://drive.google.com/drive/folders/0B7CbtsNKIEZcdIVfSlp4a0FyNHM?usp=sharing>

ID	Date	Title	How collected	Type of data
A1	Summ-13	PSQM self-evaluation	PSQM submission	Documentation
A2	Jun-13	TAPS application	emailed by school	Written tasks (researcher-led)
A3	Oct-13	Cluster day 1 initial discussion notes	TAPS Cluster day 1	Semi-structured (researcher-led) discussions or meetings
A4	Oct-13	Assessment exemplar	TAPS Cluster day 1	Documentation
A5	Oct-13	F/s-ive written task - handwritten	TAPS Cluster day 1	Written tasks (researcher-led)
A6	Oct-13	F/s-ive written task –SL - typed	TAPS Cluster day 1	Written tasks (researcher-led)
A7	Oct-13	F/s-ive sorting	TAPS Cluster day 1	Written tasks (researcher-led)
A8	Nov-13	Notes for each pyramid level	School visit 1	Semi-structured (researcher-led) discussions or meetings
A9	Nov-13	School visit 1 schedule inc SL interview	School visit 1	Semi-structured (researcher-led) discussions or meetings
A10	Nov-13	Science policy	School visit 1	Documentation
A11	Jan-14	Lesson Obs notes Y5 and Y6	School visit 2	Non-participant observation
A12	Jan-14	School visit 2 schedule inc Y5 and Y6 lessons	School visit 2	Non-participant observation
A13	Jan-14	Y5 marking sample1	School visit 2	Documentation
A14	Jan-14	Y5 marking sample2	School visit 2	Documentation
A15	Jan-14	Y5 space lesson plan	School visit 2	Documentation
A16	Jan-14	Y5 space work sample1	School visit 2	Documentation
A17	Jan-14	Y5 space work sample2	School visit 2	Documentation
A18	Jan-14	Y5 space work sample3	School visit 2	Documentation
A19	Jan-14	Y6 inheritance work sample	School visit 2	Documentation
A20	Jan-14	Y6 marking sample1	School visit 2	Documentation
A21	Jan-14	Y6 marking sample2	School visit 2	Documentation
A22	Spr-14	PSQM action plan	PSQM submission	Documentation
A23	Spr-14	PSQM School Devt Plan	PSQM submission	Documentation
A24	Spr-14	PSQM Principles	PSQM submission	Documentation

A25	Spr-14	PSQM SL log	PSQM submission	Documentation
A26	Spr-14	PSQM portfolio assessment slide	PSQM submission	Documentation
A27	Spr-14	PSQM A2 reflections on Principles	PSQM submission	Documentation
A28	Spr-14	PSQM A3 reflections on SDP	PSQM submission	Documentation
A29	Spr-14	PSQM A5 reflections on monitoring	PSQM submission	Documentation
A30	Spr-14	PSQM B1 reflections on CPD	PSQM submission	Documentation
A31	Spr-14	PSQM C2 reflections on asst	PSQM submission	Documentation
A32	Spr-14	PSQM Sect E reflections	PSQM submission	Documentation
A33	Feb-14	Pyramid self-evaluation	TAPS Cluster day 2	Written tasks (researcher-led)
A34	Mar-14	Pyramid discussion with Head	School visit 3	Semi-structured (researcher-led) discussions or meetings
A35	Mar-14	School visit 3 notes inc Head interview	School visit 3	Semi-structured (researcher-led) discussions or meetings
A36	Mar-14	School visit 3 schedule inc Y4 lesson	School visit 3	Non-participant observation
A37	Jun-14	Request for TAPS 2nd year	TAPS Cluster day 3	Written tasks (researcher-led)
A38	Jul-14	Key Qs on magnets1	School visit 4	Documentation
A39	Jul-14	Key Qs on magnets2	School visit 4	Documentation
A40	Jul-14	Pupil conferencing notes	School visit 4	Non-participant observation
A41	Jul-14	Y1 overview sheet	School visit 4	Documentation
A42	Jul-14	Y1 Skills wheel	School visit 4	Documentation
A43	Jul-14	Y1 tracking sheet	School visit 4	Documentation
A44	Oct-14	Cluster day 4 ppt extract	TAPS Cluster day 4	Documentation
A45	Nov-14	Lesson obs notes for task Y1 Dropping	School visit 5	Non-participant observation
A46	Nov-14	Lesson obs notes for task Y3 rocket balloons	School visit 5	Non-participant observation
A47	Nov-14	Lesson obs notes for task Y5 Dissolving	School visit 5	Non-participant observation
A48	Nov-14	Pupil tips for science skills	School visit 5	Documentation
A49	Nov-14	Science Qs display	School visit 5	Documentation
A50	Nov-14	Science stars new display	School visit 5	Documentation
A51	Nov-14	Y5/6 skills wheel	School visit 5	Documentation
A52	Jan-15	Cluster day 5 notes	TAPS Cluster day 5	Semi-structured (researcher-led) discussions or meetings
A53	Jan-15	exploded pyramid	TAPS Cluster day 5	Written tasks

		annotations		(researcher-led)
A54	Spr-15	P4 Peer assessment	Draft TAPS exemplar	Documentation
A55	Spr-15	P5 Act on feedback	Draft TAPS exemplar	Documentation
A56	Spr-15	T3 Floorbooks	Draft TAPS exemplar	Documentation
A57	Spr-15	M2 moderation staff meetings	Draft TAPS exemplar	Documentation
A58	Spr-15	M4 science stars displayed	Draft TAPS exemplar	Documentation
A59	Spr-15	S1 skills wheel summary of class	Draft TAPS exemplar	Documentation
A60	Spr-15	S3 summative TA range of info	Draft TAPS exemplar	Documentation
A61	Spr-15	W Principles of assessment	Draft TAPS exemplar	Documentation
A62	Jun-15	Cluster day 6 questionnaire	TAPS Cluster day 6	Written tasks (researcher-led)
A63	Jun-15	SL AfL key points	SL presentation at CPD event	Documentation
A64	Jun-15	SL CPD talk1	SL presentation at CPD event	Non-participant observation
A65	Jun-15	SL CPD talk2	SL presentation at CPD event	Non-participant observation
A66	Jun-15	SL CPD handout1	SL presentation at CPD event	Documentation
A67	Jun-15	SL CPD handout2	SL presentation at CPD event	Documentation

Points to note:

School visit 1 was made by a different researcher (Interview schedule Appendix 3F).

Many of the later school visits and observations were focused on developing assessment tasks for TAPS, so are less pertinent to this study.

Appendix 5B: School A case study codes

Origin of codes

A list of possible codes was made in advance using the RQ and TAPS pyramid boxes, these were then added to ATLAS.ti as they were found in the data. Those highlighted were used, the unhighlighted codes did not feature in this dataset so were not used. Other themes arising from the data led to the creation of new codes.

Possible codes from RQ	formative, summative, purpose,
Possible codes from TAPS pyramid	Pupil role: elicitation, LO, self asst, peer asst, act on fdbk, next steps Teacher role: plan, Questioning/discussion, obs, recording/evidence, adapting, marking, fdbk range of act, shared und, moderation, criteria, recordkeeping summarise, reports, tracking
Codes arising from data	Strategies, confidence, consistency, portfolio, PSQM, TAPS, Subject Leader role, structures

Code	Notes	Frequency
confidence	links to SL role, mostly SL supporting confidence of staff	9
consistency	is this the same as shared und (or H's standardised), is it based on an aim for reliability?	9
formative	didn't use this at the start, tried to record just as strategies bec could be f/s-ive, but SL sometimes explicit discn about f-ive	14
moderation	repeated discussion of 10 min staff mtg moderations and devt of portfolio	30
next steps	clear formative purpose here	5
portfolio	imp that did not do an updated version as planned?	9
PSQM	not seen as so much of a driver as in H, more about recognition	2
purpose	more general discussion, stepping back from detail to explore why	11
Q/discn	key theme hidden in strategies - inc pupil TALK/discussion?	19
recording	inc 'evidence', but not expressed in this way by the SL	21
records	class teacher records, those passed to next teacher	8
reporting	to parents, to SLT	3
shared und	mainly about WS stars, also scheme with key Qs	26
SL role	support less experienced staff	9
SL tracking	SL tracking?	17
strategies	formative - Qs etc	29
summative	levelling	17
TAPS	mention of TAPS	11
	Codes added when DBR development phase data added:	
self/peer asst	split from strategies	22
feedback	split from strategies	6
criteria	split from shared und- eg sci stars	22
structures	explicit school structures: science stars, scheme of work	20

Appendix 5C: School A case study codes organised by TAPS pyramid layers

I explored a number of ways to structure the case study:

- Organisation into DBR phases emphasised data collection over time, but this did not help to explain the processes of ‘formative to summative’
- Organisation into RQs led to considering formative and summative separately, but it was the relationship between the two which was of interest.
- Separation into individual codes or TAPS pyramid boxes was too atomistic and led to repetition.
- Separation into TAPS pyramid layers provided a compromise between grouping into themes, but not separating into individual codes. There is still overlap/repetition, but this organisation produced the clearest structure for the chapter.

The codes were grouped into the TAPS pyramid layers. Themes for discussion arose from the significance of individual codes and from the consideration of the grouped codes for each layer.

Pyramid layer	Code	Notes	Themes/Qs arising	
Pupil	strategies	Pupil talk – key strategy	Does pupil talk/self/peer asst feed into summative?	
Pupil	self/peer asst	Inc strategies in lessons		
Pupil	next steps	Who decides these?		
Pupil/Teacher	formative	Formative purpose	Key Qs for Ts to ask, built into planning	
Teacher	Q/discn	Key Qs for Ts to ask, built into planning		
Teacher	recording	Time?		
Teacher	fdbk	Inc marking		
Teacher	criteria	Explicit criteria		
Monitoring	moderation	Repeated mention, time commitment		Imp of explicit school structures for shared und
Monitoring	portfolio	Use in process or afterwards?		
Monitoring	records	Who for? Just numbers passed on?		
Monitoring	structures	Explicit school structures: science stars, SoW		
Monitoring	shared und	Key addition – school’s influence on pyramid?		
Monitoring	SL role	Supporting new staff		
Monitoring	confidence	Imp of SL		
Monitoring	SL tracking	Separate for kn/skills		
Summative	reporting		What is ‘best fit’? Separate systems for skills/kn?	
Summative	summative	Repeated mention of best fit		
Summative	consistency	Importance of school structures		
Whole sch processes	purpose	Formative more important?	More positive about formative?	
Whole sch processes	PSQM	As a driver?	Chapter 7	
Whole sch processes	TAPS	Impact of process on school? Impact of school on TAPS?	Chapter 7	

Chapter 6 Appendices

Appendix 6A: School B case record

The data for School B was collected March 2013 - June 2016 and included 8 TAPS cluster days, 6 school visits and two PSQM submissions.

Link to anonymised case record folder containing each item:

<https://drive.google.com/drive/folders/0B7CbtsNKIEZccHNMNk5nMUxYa1E?usp=sharing>

	Dates	Data identifier
DBR Phase 1 Exploration	March13 – Nov13	B1-Ph1 to B21-Ph1
DBR Phase 2 Development	Feb14 – Jan15	B22-Ph2 to B53-Ph2
DBR Phase 3 Implementation	March15 – June16	B54-Ph3 to B86-Ph3

ID	Date	Title	How collected	Type of data
B1-Ph1	Mar13	PSQM C2 reflection	PSQM submission	Documentation
B2-Ph1	Mar13	PSQM E reflection	PSQM submission	Documentation
B3-Ph1	June13	TAPS application	emailed by school	Written tasks (researcher-led)
B4-Ph1	Oct13	F S Qu hwr scan	TAPS Cluster day 1	Written tasks (researcher-led)
B5-Ph1	Oct13	F S Qu typed	TAPS Cluster day 1	Written tasks (researcher-led)
B6-Ph1	Oct13	F S sorting and observer notes	TAPS Cluster day 1	Non-participant observation
B7-Ph1	Oct13	Notes from cluster day 1	TAPS Cluster day 1	Non-participant observation
B8-Ph1	Nov13	TAPS school visit 1	School visit 1	Non-participant observation
B9-Ph1	Nov13	Science policy	School visit 1	Documentation
B10-Ph1	Nov13	Interview Notes with SL	School visit 1	Semi-structured (researcher-led) discussions or meetings
B11-Ph1	Nov13	Planning and work egs Y1	School visit 1	Documentation
B12-Ph1	Nov13	Planning and work egs Y2	School visit 1	Documentation
B13-Ph1	Nov13	Planning and work egs Y34	School visit 1	Documentation
B14-Ph1	Nov13	Planning and work egs Y56	School visit 1	Documentation
B15-Ph1	Nov13	Y6 marking egs	School visit 1	Documentation
B16-Ph1	Nov13	Extra planning eg	School visit 1	Documentation
B17-Ph1	Nov13	Materials to support planning	School visit 1	Documentation
B18-Ph1	Nov13	Doing sci diff shts	School visit 1	Documentation
B19-Ph1	Nov13	Tracking and report egs	School visit 1	Documentation
B20-Ph1	Nov13	APP grid	School visit 1	Documentation
B21-Ph1	Nov13	Individual levels grid	School visit 1	Documentation
B22-Ph2	Feb14	School visit 2 Y34 Y56	School visit 2	Non-participant

				observation
B23-Ph2	Feb14	Y34 planning for obs lesson	School visit 2	Documentation
B24-Ph2	Feb14	Y34 Display ch ideas1	School visit 2	Documentation
B25-Ph2	Feb14	Y34 Display ch ideas2	School visit 2	Documentation
B26-Ph2	Feb14	Y34 Rtn to thought shower	School visit 2	Documentation
B27-Ph2	Feb14	Y34 Pupil self asst	School visit 2	Documentation
B28-Ph2	Feb14	Y34 multiple choice	School visit 2	Documentation
B29-Ph2	Feb14	Y56 planning for obs lesson	School visit 2	Documentation
B30-Ph2	Feb14	Y56 lesson notes	School visit 2	Non-participant observation
B31-Ph2	Feb14	Y56 rev and irrev elicitation	School visit 2	Documentation
B32-Ph2	Feb14	Y56 differentiated table1	School visit 2	Documentation
B33-Ph2	Feb14	Y56 differentiated table2	School visit 2	Documentation
B34-Ph2	Feb14	Y56 marking	School visit 2	Documentation
B35-Ph2	Feb14	Y56 targets	School visit 2	Documentation
B36-Ph2	Feb14	1st pyramid self evaln	TAPS Cluster day 2	Written tasks (researcher-led)
B37-Ph2	Mar14	Head interview	School visit 3	Semi-structured (researcher-led) discussions or meetings
B38-Ph2	Mar14	Head pyramid discn	School visit 3	Semi-structured (researcher-led) discussions or meetings
B39-Ph2	Apr14	PSQM self asst	PSQM submission	Documentation
B40-Ph2	June14	Moderation mtg notes	School visit 4	Non-participant observation
B41-Ph2	June14	Modn mtg pupil work1	School visit 4	Documentation
B42-Ph2	June14	Modn mtg pupil work2	School visit 4	Documentation
B43-Ph2	June14	Modn staff mtg agenda	School visit 4	Documentation
B44-Ph2	June14	Modn mtg level child	School visit 4	Documentation
B45-Ph2	June14	Modn mtg how level	School visit 4	Documentation
B46-Ph2	June14	Request for TAPS 2nd yr	TAPS Cluster day 3	Written tasks (researcher-led)
B47-Ph2	Oct14	Checking pyramid eggs	TAPS Cluster day 4	Written tasks (researcher-led)
B48-Ph2	Oct14	Target eggs	School visit 5	Documentation
B49-Ph2	Nov14	Nappies lesson	School visit 5	Non-participant observation
B50-Ph2	Nov14	Wonder wall	School visit 5	Documentation
B51-Ph2	Nov14	Y56 diff HWK	School visit 5	Documentation
B52-Ph2	Jan15	Cluster day 5 notes	TAPS Cluster day 5	Non-participant observation
B53-Ph2	Jan15	Pyramid annotated	TAPS Cluster day 5	Written tasks (researcher-led)
B54-Ph3	Mar15	PSQM Action plan	PSQM submission	Documentation
B55-Ph3	Mar15	PSQM C2 refln	PSQM submission	Documentation
B56-Ph3	Mar15	PSQM SDP	PSQM submission	Documentation

B57-Ph3	Mar15	PSQM CPD	PSQM submission	Documentation
B58-Ph3	Mar15	PSQM Section E refln	PSQM submission	Documentation
B59-Ph3	Mar15	PSQM portfolio	PSQM submission	Documentation
B60-Ph3	Mar15	PSQM SL log	PSQM submission	Documentation
B61-Ph3	Mar15	Eg group elicitation	Draft TAPS example	Documentation
B62-Ph3	Mar15	Eg Identifying next steps	Draft TAPS example	Documentation
B63-Ph3	Mar15	Eg adapt challenge	Draft TAPS example	Documentation
B64-Ph3	Mar15	Eg modn staff meeting	Draft TAPS example	Documentation
B65-Ph3	Mar15	Eg struct for record kping	Draft TAPS example	Documentation
B66-Ph3	Mar15	Eg Hwk	Draft TAPS example	Documentation
B67-Ph3	May15	Eg new sci skills toolkit	Draft TAPS example	Documentation
B68-Ph3	May15	Y1 recording	School visit 6	Documentation
B69-Ph3	May15	Y1 recording2	School visit 6	Documentation
B70-Ph3	May15	Y1 recording3	School visit 6	Documentation
B71-Ph3	May15	Y1 recording4	School visit 6	Documentation
B72-Ph3	May15	Y1 recording5	School visit 6	Documentation
B73-Ph3	May15	Y1 recording6	School visit 6	Documentation
B74-Ph3	Jun15	Impact Q Hwr	TAPS Cluster day 6	Written tasks (researcher-led)
B75-Ph3	Jun15	Cluster day 6 notes	TAPS Cluster day 6	Non-participant observation
B76-Ph3	Jun15	Impact Qu typed	TAPS Cluster day 6	Written tasks (researcher-led)
B77-Ph3	Nov15	Impact survey hwr	TAPS Cluster day 7	Written tasks (researcher-led)
B78-Ph3	Nov15	Impact survey typed	TAPS Cluster day 7	Written tasks (researcher-led)
B79-Ph3	Nov15	Pyramid self eval	TAPS Cluster day 7	Written tasks (researcher-led)
B80-Ph3	May16	SL Presn planning	SL presentation	Documentation
B81-Ph3	May16	SL Presn ppt	SL presentation	Documentation
B82-Ph3	May16	SL Presn notes	SL presentation	Non-participant observation
B83-Ph3	Jun16	SL interview	TAPS Cluster day 8	Semi-structured (researcher-led) discussions or meetings
B84-Ph3	Jun16	SL TAPS day slide	TAPS Cluster day 8	Documentation
B85-Ph3	Jun16	SL presn transcription	TAPS Cluster day 8	Non-participant observation
B86-Ph3	Jun16	Ts next steps	TAPS Cluster day 8	Written tasks (researcher-led)

Points to note:

School visit 1 was made by a different researcher (interview schedule Appendix 3F).

Many of the later school visits and observations were focused on developing assessment tasks for TAPS, so are less pertinent to this study.

Appendix 6B: School B case study codes

After coding each DBR Phase, the frequency of codes was noted to support analysis of themes.

Once coding was complete, possible patterns/themes were coloured:

Areas of	changing focus	consistent focus for devt	possible pattern/point to note
----------	----------------	---------------------------	--------------------------------

	Docs	B1-21	B22-53	B54-86	
Codes	Notes	Phase1	Phase 2	Phase3	Total
challenges	barriers to developing asst	1	2	2	5
confidence	confidence of Ts	1	1	11	13
consistency	key area for SL, how to be consistent in sch processes, linked to reliability?	12		3	15
criteria	use a number of docs to support judgements	10	21	11	42
differentiation	diff act/Qs planned for diff gps, are ch set? Expectations set?	4	21	3	28
evidence	children's recording	5	9	2	16
formative	identifying where learner is/go next	8		12	20
knowledge	concepts, vocab, invest in context, prob for mixed age classes?	6	5	2	13
levelling	key part of consistency for SL, levelness	15	14	3	32
marking	split from strategies in Phase 2 analysis	14	10	4	28
moderation	recent devt, discuss with staff	8	11	10	29
next steps	for children, inc targets - esp in Phase 2 analysis	7	5		12
parents	added in Phase 2 analysis	2	4	7	13
planning	split from records in Phase 2 analysis	2	10	6	18
PSQM	key driver to start self-evaluation process	5		7	12
purpose	purpose of asst defined by SL	2		2	4
records	teacher records to be passed on, inc planning in Ph2	4	8	2	14
reporting	to parents, Head or Governors	3	2		5
Self-evaluation	SL evaluating school practice inc TAPS	11		12	23
self/peer asst	updated from 'active pupils' when started Ph2	6	19	24	49
sharing practice	sharing strategies with other staff in sch (split from staff team)	3	1	8	12
skills	sci enq or Working Scientifically	10	11	7	28
staff team	imp of all working tog on this, inc shared und	5	7	10	22
strategies	for formative asst inc T Q, marking?	7	22	19	48
structures	docs for criteria or processes - lots of these used	16	7	7	30
summative	largely linked with levelling during Phase 2 analysis	17	4	13	34
TAPS	added in Phase 3 analysis			28	28
tracking	summarising and passing info to Head, SL or next T	12	2	5	19

Appendix 6C: Themes

During and after coding themes arising from the data were noted. In order to structure these for the chapter, the codes/themes were sorted by TAPS pyramid layer.

RQ focus: change over time, relationship between formative and summative.

Ch	Pyramid layer	Themes arising	Codes	Key Changes
6.2	Whole sch processes	rel btw form and sum – beg, end, changes?	Purpose, formative, summative	Formative to sum?
6.3	Summ reporting	where does the info go	Confidence, tracking, reporting, parents	What to track?
	Summ reporting	changing importance	evidence	changing importance
6.4	Monitoring	concern to use same structures for consistency, increasing list	Consistency, structures	Search for consistency
	Monitoring	focus for summ, level child/wk	Levelling, criteria	Dev rel
	Monitoring	to support levelling/expectations	moderation	Expl judgements
6.5	Teacher	from closed to open?	differentiation	More open
	Teacher	annotating planning - for who? Inc levelling and space for notes	planning	Annotating planning
	Teacher	focus on marking but man? Time for pupils to respond	marking	Less of a focus
6.6	Pupil	strategies for formative asst, range trialled and evaluated rel btw next steps and targets	Strategies next steps	Tried new strat
	Pupil	key devt area identified by sch <i>ch role of T if pupils more active?</i>	self/peer asst	Ch more active take role of T

Appendix 6D: Card Sort

Strategies listed from the PSQM Round 4 data, for sorting with formative and summative in mind.

Self assessment	Observation	Tests
Peer assessment	Photographs	Quiz
Marking	Group challenge	Mind map
Tracking grid	Pupil presentation	KWL grid (<u>K</u> now, <u>W</u> ould like to, <u>L</u> earnt)
Expectations on planning	Investigation	Concept cartoon
Success criteria	Pair talk	Pupils pose questions
Best work folder	Discussion	
Post its	Questioning	

Appendix 6E: List of conferences/teacher presentations (Jan 14 – July 16)

* = used formative/summative card sort activity

Date	Organisation	Audience
Jan14	ASE Conference Birmingham *	20 teachers and advisors
March14	National Science Centre, York *	12 science subject leaders
April14	Kirklees LA *	50 science subject leaders
June14	Bristol network *	40 science subject leaders
June14	Camden *	60 primary and secondary teachers
June14	PSTT College Conf, Manchester *	30 PSTT College Fellows/Trustees
Sep14	BERA	Early career researcher plus main conference session
Oct14	PSTT Conference	Report launch to whole conference, more detail for 40 delegates in workshop
Oct14	Stockport subject leader *	45 teachers and 3 trainers
Nov14	ASE region conference at Bath Spa *	Report launched to whole conference (130), 20 delegates in workshop
Nov14	Wellcome mtg primary science 'expert' group	reports shared with researchers, teachers and subject associations (30)
Jan15	ASE Conference Reading	70 delegates across 2 sessions
Jan15	PSTT Hubs mtg	TAPS reports shared with other Hubs
Feb15	Cardiff ASE Teachmeet	30 teachers/subject leaders
Feb15	Bristol ASE Teachmeet	30 teachers/subject leaders
Feb15	S Glos Best Practice Forum	20 subject leaders
March15	Trowbridge Schools collaborative	15 subject leaders
March15	BSU PGCE	300 trainee teachers
March15	Teach First	15 trainee teachers
April15	PSTT Fellows, Hubs and Learned societies	Updated report emailed out to all: 150 fellows, research hubs at 5 other universities, societies
May15	BSU Ed Studies	100 Education Studies 2nd yr students
May15	Wellcome mtg	TAPS reports shared with 50 primary science community representatives
May15	Guernsey	12 subject leaders
June15	Bristol Subject Leaders	30 subject leaders
June15	School Direct	100 trainee teachers
June15	Salisbury	15 subject leaders and PSTT Southampton
June15	Oxford Brookes	100 teachers/subject leaders plus PSTT Oxford Brookes and Science Oxford
Sep15	ESERA conference, Helsinki	40 educators in TAPS session, poster and proceedings for international audience of 1300
Oct15	PSQM conference	100 PSQM Hubleaders from across UK

Oct15	PSTT conference	100 PSTT college fellows, trustees, ACs
Nov15	Dorchester conference *	30 science subject leaders (SSL)
Nov15	Glos network series x3 * 1 x Coleford 2 x Cheltenham	30 SSL in Coleford 60 SSL in Cheltenham day 1 60 SSL in Cheltenham day 2
Nov15	ASE West conference at Bath Spa	(2 sessions by led TAPS teachers) 20 teachers/advisors x2
Nov15	S Glos network	30 science subject leaders
Dec15	Engage 15 – poster at public engagement conf	200 delegates - range of public engagement professionals from universities and charities
Jan16	ASE National conference, Birmingham	30 advisors and teachers in sessions, approx. 200+ booklets handed out on PSTT stands over 3 day conference
Jan16	Melksham cluster	10 science subject leaders
Feb16	Berkshire conference	30 science subject leaders
Feb16	ASE Teachmeet, Bristol	20 pri and sec
Feb16	PGCE Primary	300 trainee teachers
March16	Surrey conference	40 SSL
March16	Teachfirst	20 trainees
March16	Bristol School Direct	20 trainees
May16	Trowbridge cluster	10 subject leaders
May16	NSLC York	10 teachers in session, 150 bklets in bags
May16	Sheffield Learning First	500 Ts and educators, plus filmed for youtube
June 16	PSTT Belfast	35 am session, 20 pm session, 350 bklets in bags
June 16	TAPS celebration at Bath Spa	60+ join in pm - local schools plus Welsh schools and Cardiff Met
July16	ASE Summer conference, Hatfield	20 advisors/ITE in session, 50 made aware of TAPS at conference
July16	PSQM Senior Hubleaders meeting	12 at meeting - responsible for areas of UK, planning how to intro to their hubleaders
July 16	Devon conference	50 science subject leaders