International Journal of Science Education

RESEARCH ARTICLE

Development and Exemplification of a Model for Teacher Assessment in Primary Science

D. J. Davies[1*], S. Earle[2], K. McMahon[2], A. Howe[2] & C.Collier[2] [where * denotes corresponding author, see below]

[1]*Cardiff Metropolitan University, UK;* [2]*Bath Spa University, UK*

Abstract

The Teacher Assessment in Primary Science (TAPS) project is funded by the Primary Science Teaching Trust (PSTT) and based at Bath Spa University. The study aims to develop a whole-school model of valid, reliable and manageable teacher assessment to inform practice and make a positive impact on primary-aged children's learning in science. The model is based on a data flow 'pyramid' (analogous to the flow of energy through an ecosystem) whereby the rich formative assessment evidence gathered in the classroom is summarised for monitoring, reporting and evaluation purposes (Nuffield Foundation, 2012). Using a Design-Based Research (DBR) methodology, the authors worked in collaboration with teachers from project schools and other expert groups to refine, elaborate, validate and operationalise the data flow 'pyramid' model, resulting in the development of a whole-school self-evaluation tool. In this paper we argue that a DBR approach to theory-building and school improvement drawing upon teacher expertise has led to the identification, adaptation and successful scaling-up of a promising approach to school self-evaluation in relation to assessment in science.

Keywords*: Primary science education; Teacher assessment; Formative assessment; Design-Based Research*

Introduction

Assessment of primary science in England since 2010 has shifted away from national testing towards teacher assessment, reflecting a growing awareness in the teaching profession and educational research community of the harmful effects of high-stakes summative testing (Newton, 2009) together with its distorting effects on the taught curriculum (Wiliam, 2003). Arguably this aligns practice with that in other countries that perform well in science education (e.g. Finland, Australia, Estonia), where assessment systems utilise teachers' professional judgement of pupil attainment and progress, based on evidence gathered through everyday classroom work (PISA, 2012; McIntyre, 2015). In England this change may also reflect a decline in the status of science as a 'core' subject in the primary national curriculum (Boyle & Bragg, 2005; Blank 2008), since the other core subjects of English and mathematics continued to be externally tested.

*Corresponding author. Email: djdavies@cardiffmet.ac.uk

A further change occurred in 2013 with the proposal in a revised national curriculum (DfE, 2013) to

discontinue the pupil attainment levels originally established by the Task Group on Assessment and Testing (TGAT, 1988). This signalled an apparent shift to greater freedom for primary schools to develop their own approaches to assessment (DfE, 2014). However, external accountability through inspection remained strong, with clear views on the value of internal testing being expressed by the Chief Inspector:

> We will not endorse any particular approach. But we do expect every school to be able to show what their pupils know, understand and can do through continuous assessment and summative tests. (Sir Michael Wilshaw's Speech at the North of England Education Conference in January 2014)

The perceived tension between new freedoms and concern about hidden expectations for external accountability have led to considerable diversity in the way schools responded to these assessment changes (Riddell, 2016). The Teacher Assessment in Primary Science (TAPS) project aimed to provide greater clarity in this situation by developing a whole-school model of valid, reliable and manageable teacher assessment, to inform practice and make a positive impact on primary-aged children's learning in science.

An increasing reliance on teacher assessment raises questions about whether evidence of pupil learning collected for the formative purposes of supporting learning can legitimately be used to summarise attainment against external criteria. Harlen (2013) asserts that any assessment opportunity can be used for both formative and summative purposes. The 'day-to-day, often informal, assessments' (Mansell et al, 2009, p. 9), which are used to inform next steps in learning, can also be summarised at a later date, whilst, conversely the results from summative tests can be used formatively to guide learning (Black et al 2003). However, Gipps and Murphy (1994, p. 14) argue that 'any attempt to use formative assessment for summative purposes will impair its formative role', since they constitute two 'paradigms' of assessment (Gipps 1994) - what Stiggins (1992) has referred to as 'trickle up' and 'trickle down' respectively - whose concerns are fundamentally different; the former concerned with classroom decision-making whilst the latter emphasises standardisation and accountability. This view has been challenged by Wiliam and Black (1996), who argue that the elicitation of classroom-based assessment evidence can serve both purposes if its collection is separated from its interpretation or judgement. Taras (2007, p. 367) distinguishes between the processes of assessment and its functions, which may be 'multifarious' but do not impinge on the processes (Taras, 2009, p. 59). She argues that 'fear of misuse' of judgements has resulted in a gradual separation of the formative and summative functions into Gipps' paradigms, whereas if process is emphasised the dichotomy disappears, such that formative assessment can be regarded as summative assessment with feedback.

In order to be fit for the purposes of enhancing, recording and reporting on learning, educational assessment needs to meet certain criteria for quality, usually listed as validity, reliability, manageability and impact (e.g. Harlen, 2013). Validity concerns whether the assessment is able to capture learning for the purposes to which the resulting data are to be put. Gardner et al. (2010) argue that teacher assessment has greater validity than testing because it can be based on a wider range of types of evidence, collected in a range of contexts. This is particularly relevant for assessment of practical and collaborative inquiry-based science education (IBSE) which develops skills - such as forming hypotheses - and attitudes - such as respect for evidence - that are not adequately examined in a test (Harlen & Qualter, 2014). However, Stobart (2012) asserts that validity is intrinsically linked to purpose, which raises the question of whether validity is compromised if the same evidence is used for both formative and summative purposes. Validity may take different forms, for example content or construct validity concerns the extent to which a particular assessment instrument represents the range of skills and understanding for a particular topic area. Construct under-representation is a threat to

validity, especially in primary science where the key skills of scientific inquiry may be more difficult to assess than conceptual understanding in classroom environments – though arguably more easily than in a test. Construct irrelevance is also a danger for all teacher assessment; Johnson (2013, p. 99) found teachers to be consciously or unconsciously influenced by construct-irrelevant pupil characteristics (e.g. gender, ethnicity, socioeconomic status). Other relevant forms of validity include participant-confirmed validity, whereby the pupil recognises themselves in the judgement of the teacher – highlighting the role of self and peer-assessment (Wiliam 2011) – and consequential validity, concerning the extent to which information gathered for formative purposes is used to support pupil learning. This form of validity could also be termed *impact*, since a criterion for effectiveness in assessment should be the benefit it confers to learners, by contrast with the potential negative impact (in the form of examination anxiety and boredom from revision) of external testing.

The reliability of an assessment – broadly definable as the extent to which the results can be said to be of acceptable consistency or accuracy for a particular use (Harlen 2013, p. 9) - can also take a variety of forms, including inter-rater reliability, which concerns whether the same judgement would be made on the same evidence by different assessors. This is a potential weakness of teacher assessment (Black et al, p. 2011), particularly when it involves rating annotated samples of pupil work against external criteria (Klenowski & Wyatt-Smith, 2010). External testing using standardised instruments can be argued to produce results of greater consistency, whose reliability is measurable. In England, the discontinuation of national testing in science arguably removed its function in creating a shared view of a level of achievement. Whilst few studies have attempted to assign coefficients of reliability to teacher judgements, there is evidence that it can be improved through consensus moderation (Johnson, 2013) involving discussion of samples of pupil work against criteria. In England the reduced status of science (relative to English and maths) has meant there is currently little support for moderation at national and local level. Some jurisdictions such as Queensland, Australia employ external moderation and exemplification of criteria to support teachers' judgements (Klenowski, 2011), though studies of moderation processes have found that it takes up to three years to achieve acceptable inter-rater reliability through such approaches (Stanley et al, 2009). Wiliam (2003) argues that, whilst teacher assessment can become more reliable, there is inevitably a 'trade off' between reliability and validity since the wider range of evidence required to represent the constructs within a field of learning may be more difficult to rate consistently on an external scale than the relatively narrow dataset obtained through a single assessment instrument.

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

Origins of the data-flow pyramid model

The model of whole-school primary science assessment upon which the TAPS project is based started as a visual representation of a 'framework for assessment of science in primary schools' proposed by a working group of science education experts convened by the Nuffield Foundation (2012) under the leadership of Professor Wynne Harlen. The group argued that pupils' ability to 'work scientifically' - planning and carrying out enquiries and applying their knowledge in new contexts through discussion - is best assessed in the context of these activities rather than through written tests. Recognising the need for a coherent approach - in which the reporting of summaries of what has been learned supports that learning - their aim was to develop an assessment framework that:

> … sets out how evidence of pupils' attainment should be collected, recorded, communicated and used by those involved in pupils' education. It describes how the dependability of the resulting information can be optimised for different purposes and what support is needed to implement the procedures (Nuffield Foundation, 2012, p. 9).

The resulting framework divided science assessment purposes and processes into those concerned with individual pupils (both formative and summative); class and school records; and sample test data which would be used for evaluating national performance. To illustrate the flow of assessment data from individual to whole-school diagrammatically, the working group chose as an analogy the flow of energy through a biological ecosystem, with the various purposes for the assessment (formative, recording, summative, reporting) conceived as analogous to trophic levels within a 'pyramid of numbers' (Figure 1): The use of analogies to help explain science concepts is common practice (Coll et al, 2005), though the application to an aspect of professional practice appears to be novel.

Whilst at the 'ongoing formative assessment' level at the base of the pyramid in Figure 1, a wide range of evidence would be used to inform teaching and feedback to learners, only a proportion of relevant data would feed up to the next level ('annual reporting to parents') and successively smaller selections of assessment information would flow upwards to the more summative levels of the pyramid.

> Thus there is a gradual reduction in the breadth and detail of information that is recorded and reported, from the rich formative assessment to the succinct, summative information (Nuffield Foundation, 2012, p. 19).

Key to this model is the change of function of formative assessment data for summative purposes. Although how and where this change of function takes place is not explicit in Figure 1, the authors recommend that: 'the translation of detailed formative data to summative judgements should be moderated within the school, using group procedures and reference to national exemplars.' (Nuffield Foundation 2012, p. 8). The recommendation that this moderation should take place at the end of 'key stages' (at ages 7, 9 and 11) implies that it occurs at the transition between levels three and four in Figure 1, when information about individual pupils in the first three layers is aggregated to become information about groups or cohorts of pupils in the fourth and fifth layers. Moderation of teacher judgements is required for individual pupils as well as whole classes, so the decision to divide the assessment framework by scale (individual/class/school) and time (ongoing/year/Key Stage) becomes problematic in mapping the relationship between formative and summative purposes of assessment.

This is one of the issues which the TAPS project sought to address in translating Figure 1 into a form which primary schools could use to examine critically their own use of teacher assessment in science. Since the principal audience for the 2012 report was policymakers, the level of detail and exemplification relating to

classroom practice required to 'operationalise' Figure 1 for school use was not yet present. The project sought to develop this data-flow pyramid model to fulfil two purposes:

1. As a theoretical model of how a whole-school system for the collection, feedback and summary of pupils' science learning assessment data for formative and summative purposes could fulfil the quality criteria of validity, reliability, manageability and impact.
2. As a whole-school self-evaluation tool to be used by science subject leaders and others to identify strengths and weaknesses in primary teacher assessment of science and thereby to plan for enhanced quality in this aspect of professional practice.

Whilst purpose 1 above could be regarded as primarily academic and purpose 2 professional in focus, we would argue that the former underpins the latter. Whilst a self-evaluation tool could be used instrumentally as a check-list of features that need to be present to ensure quality, using the model as a means of developing a conceptual understanding of the fundamental principles of assessment – what Klenowski (2011) has described as 'assessment literacy' – potentially increases its power for teachers' professional development.

Methodology

We decided to adopt a Design-Based Research (DBR) methodology (Brown, 1992) since this approach aims to engineer products and develop recommendations which will support educational reform and inform practice, addressing concerns about the lack of impact of educational research on school and classroom practice (Hartas, 2010). In DBR the development of theory and products to support practice are intertwined; our design goal was to develop Figure 1 into both a theoretical model and a self-evaluation tool which would have practical impact, by explaining and exemplifying what such a system would look like in practice, adapting and adding to the model in response to user feedback whilst maintaining a focus on validity, reliability and manageability. Since the aim of DBR is to 'generate evidence-based and ecologically-valid recommendations for practice' (McGuigan & Russell 2015, p. 35), the approach necessitates a collaboration between researchers and practitioners in real contexts (Anderson and Shattuck, 2012). The continuous cycles of designing and testing within DBR require theories to do 'real work' (Cobb et al, 2003); thus the TAPS project schools used the data-flow pyramid model as a self-evaluation tool from the outset, which then went through a number of versions as the design principles evolved (Anderson and Shattuck 2012). The phases of this iterative development process are summarised in Table 1.

[insert table 1]

Applications to participate in TAPS were invited from primary schools in South West England and - from the 50 applications - 12 were selected to represent a range of size, locality and approaches to teacher assessment in science, together with a commitment to develop in this area. The project alternated cluster meetings (three per year) and school visits (n=72) where a range of qualitative data was collected, including school assessment policies, records and other documentation; classroom observations of teachers carrying out science assessment in Years 1 to 6 (ages 5 to 11); observations of staff meetings and moderation sessions; interviews with science subject leaders and assessment co-ordinators; collection of annotated samples of pupil science work; participant validation questionnaires; records of discussion and teacher annotations on successive versions of the data-flow pyramid during cluster days. Data were analysed thematically, using the developing categories within the pyramid as an analytical framework. Thus, for example, notes from lesson observations were coded against statements drawn from Harlen (2013 - see Table 2) related to the 'ongoing formative

assessment' layer in Figure 1.

Shavelson et al (2003) suggest that although DBR can address the complexity of interventionist studies, narrative accounts can risk circularity in their claims. In order to test the validity and reliability of the TAPS products and conclusions, the research team sought the views of a wider constituency of educators than those within the project schools. By drawing on 'expert teachers' (who had received Primary Science Teacher Awards) and 'expert schools' (those holding the Primary Science Quality Mark Award), together with external advisors, the model was validated by those who had not been involved in its production. Validation data included lesson observations of and subsequent interviews with 'expert teachers' (n=4), together with school assessment documentation and children's work, which were analysed thematically against the statements in Figure 4, to determine the weight of evidence for each element of the pyramid model and identify gaps. Focus group interview data from a validation panel of 'experts' (see phase 4 below) were similarly analysed using the online published version (Figure 5) as an analytical frame. Download statistics for this version also provide an indication of user-confirmed validity.

Findings

In keeping with the principles of DBR, the findings below are presented as a process of development of the model/tool, under the headings within Table 1 above. It should be noted here that several versions of the DBR process are described in the literature (Easterday et al., 2014); the phases defined below represent a synthesis selected for their applicability to the process undertaken.

Phase 1: Understanding and defining

At the first TAPS project cluster day in September 2013, the 36 participating teachers from 12 primary schools (in most cases the science subject leader, assessment co-ordinator and information and communications technology [ICT] co-ordinator) were asked whether the assessment framework represented in Figure 1 corresponded with practice in their own schools. Whilst there was general recognition of the relevance of the model, one of the suggestions to emerge from the subsequent discussion was the insertion of a level relating to the monitoring of pupil progress against assessment criteria between levels 1 (ongoing formative assessment) and 2 (annual reporting to parents) as it was felt that the making of judgements and summary of data would be necessary in order to make such reports. Participants agreed that, rather than only being asked for summative judgements at the end of key stages (level 3 in Figure 1) their school pupil attainment tracking systems required them to make such judgements continuously – or at least at regular intervals throughout a school year. It is at this intermediate level that the process of assessment data reduction and re-purposing for summative uses was mainly occurring in their schools. Participants were also given a copy of *Assessment & Inquiry-Based Science Education: Issues in Policy and Practice* (Harlen, 2013). In addition to outlining principles for formative and summative assessment of inquiry-based science education (IBSE) across Europe, this report proposes two sets of 'standards': for classroom assessment practice (Table 2) and for use by school management teams (Table 3):

[Insert tables 2 and 3 here]

It was agreed by the participants that the 'standards' outlined in Tables 2 and 3 constitute a useful set of descriptors of effective practice against which their schools' assessment of pupil learning in science could be

evaluated. Accordingly, when we made our first round of school visits between November and December 2013, the TAPS research team used a subset of the 'standards' in Table 2 which we deemed to be observable as a framework for classroom observation and developed a set of interview questions for science subject leaders (see Appendix 1) based on the levels of the data-flow pyramid model and informed by the 'standards' in Table 3. Analysis of data from the first round of school visits - consisting of 17 lesson observations and 11 interviews (principally with the science coordinator) - informed our understanding of the 'problem' (Bryk et al, 2010); that of the fracture between ongoing teacher assessment and end of year tests (see also Earle, 2014). This analysis, together with the discussion of Harlen's (2013) 'standards' on day 1, suggested a need to develop and support teachers' repertoire of assessment practices at all levels of the data-flow pyramid. One way of achieving this could be to insert the most pertinent 'standards' statements into the appropriate level of the pyramid as an operationalisation of the Nuffield model. Together with the insertion of a 'monitoring of student progress' level between levels 1 and 2, this synthesis resulted in the first version of the 'TAPS Project Whole-School Science Assessment Evaluation Tool' (Figure 2), which also represents the project's' first attempt at a theoretical model of teacher assessment. Figure 2 includes a 'RAG-rating' (red, amber or green) for each statement, as this is a familiar process for primary teachers, both in relation to pupil assessment records and school self-evaluation documentation. The key in Figure 2 gives some indication of what each colour implies; the intention of the tool was not to compare schools but to promote discussion within schools regarding effectiveness of assessment processes.

[Insert Figure 2]

Phase 2: Development

On cluster day 2 (February 2014) participant teachers were asked in 'within-school groups' to undertake an initial trial of the self-evaluation tool by RAG-rating each statement on the basis of their current awareness of their school assessment processes. The outcome of this exercise is summarised in Table 4 below.

[Insert table 4]

As can be seen from Table 4 above, all statements in the pyramid tool received at least five self-assessments of at least 'some evidence' from participants, suggesting that the 16 statements are broadly grounded in schools' experience. The statement for which participants felt they had most evidence was: '*Teachers base their judgements of students' learning outcomes on a range of types of activity.'* This suggests a concern for validity in the assessment process, with the acknowledgement that single sources of evidence may not adequately represent a pupil's understanding of a scientific concept or procedural capability. The statements for which participants felt their schools had the least evidence were:

> *(Reports) provide information about assessment processes.*
> *Teachers take part in discussion with each other of students' work in order to align judgements of levels or grades.*
> *Students are aware of the criteria by which their work over a period of time is judged.*

Two of these statements concern making assessment processes transparent to those they affect (pupils and their parents), whilst the other relates to increasing the reliability of teacher judgements through consensus moderation, which is a feature of 'mature' systems of teacher assessment such as Queensland (Klenowski, 2011, see above).

Development and Exemplification of a Model for Teacher Assessment in Primary Science

Uncertainty was expressed by some participants about terms used in some of the statements (e.g. 'study of products' to refer to teachers' use of samples of pupil written or other physical samples of work). There were also felt to be inconsistencies in terminology (e.g. references to 'students' and 'children'). The 'feedback arrows' to the right of the pyramid were felt to be potentially confusing, as they implied that information (in the form of evidence and teacher judgements) would only be transferred from higher to lower levels, whereas the original Nuffield Foundation model (Figure 1) had emphasised the 'upward' flow of data. Importantly, the role of pupils in their own and peer assessment was felt to be under-represented. A guest-speaker presentation earlier that day from a representative from the Association for Achievement and Improvement through Assessment (AAIA) had highlighted the value of pupil self and peer assessment, and the literature on pupil role (e.g. Wiliam, 2011) also suggested that this aspect required stronger representation in the model.

During the first round of school visits it had also became apparent that pupils' involvement in the process of assessment was under-represented in the first version of the evaluation tool. The visits helped to identify further 'standard' statements that captured the variety of ways pupils were involved. Specifically ~~it was noted that there~~ , in the 17 lesson observations were several instances of learning objectives and success criteria being shared by teachers with pupils without it being clear whether pupils had understood them. Although feedback was being given by teachers to pupils, it was unclear whether pupils had acted on this. ~~Pupils, it emerged,~~Science subject leaders reported during interviews that pupils were often involved in assessing their own ideas but that the prevalence of ~~pupils assessing peers' ideas~~peer assessment was much lower. From the observations, t~~T~~here was little evidence that pupils were involved in identifying next steps in learning. Explicitly including standard statements that related to these aspects of assessment became a focus of development following cluster day two. Taking into account school visit 1 analysis and participant feedback, the research team added a new 'level 1- pupil layer' at the base of the pyramid using statements from Short (2014) which had been presented by the speaker from AAIA. A clear 'upward' arrow was also included at the centre of the model to emphasise the predominant direction of assessment data flow. These changes are highlighted in Figure 3 below.

[Insert figure 3]

Phase 3: Exemplification

Further analysis of the participant annotations of Figure 3 involved a thematic treatment of the evidence they described to demonstrate each statement. Beyond providing insights into individual schools, this process revealed that statements were being interpreted in different ways. For example, the statement *'Teachers gather evidence of their students' learning through study of products relevant to the learning goals'* differed according to whether pupil-produced outcomes had to be written or could include a diversity of modes and whether the process or final outcome should be documented. Comments showed that teacher activity producing outcomes such as marking and subject leader scrutiny of work sample were also considered as evidence for this statement. The statement *'Teachers take part in discussion with each other of students' work in order to align judgements of levels or grades'* was understood as moderation by 6 of the 12 schools, and use of the word moderation was considered to capture the shared meaning better. However 7 of the 12 schools could not exemplify how to moderate judgments of attainment in science, with practical science enquiry emerging as an area that was difficult to moderate. The statement *'A manageable system for record keeping is in operation to track and report on students' learning',* was RAG-rated by 10 of the 12 schools as

green or amber, however two schools cited 'end of unit assessments' as evidence, indicating that that the summative judgments on individual children did not derive from data flowing through the pyramid.

It became apparent that as participants' interpretations were informed by their widely differing experience and expertise in assessment and science pedagogy, exemplification of each statement would support both interpretation of terms and the development of teachers' repertoires of assessment strategies. Discussion of data from tutor visits to different schools raised concerns that what was manageable for one school context may not be manageable for another and that no single system should be presented as exemplification. This concern that exemplification could constrain rather than support schools was echoed by expert members of the TAPS advisory board (minutes 26/3/2014, 3/2015) who argued that good exemplification should not lead to narrowing of practice, but should support a diverse and creative range of teacher responses. Accordingly, multiple examples of practice taken from school visit data were foregrounded by agreement within the project team to create short descriptions under statements within the model. For instance, to exemplify the statement: *'Pupils use assessment to advance their learning by acting on feedback',* the following text was added to the box: *'e.g. respond to mini plenary advice in second half of the lesson, make improvements in next investigation.'* Explanations of the feedback arrows to the right of the pyramid were added together with the opportunity for schools to 'RAG-rate' themselves in relation to their use of judgements to feed back into lower levels. The resulting version, with changes highlighted, is in Figure 4 below.

[Insert figure 4]

Figure 4 was presented back to participants during cluster day 3 (June 2014). The inclusion of explanatory text enabled discussion to shift from concern about the meaning of particular statements to offering examples from each school's practice to demonstrate them. To encapsulate the range within each statement, the exemplification included documents and other items produced by schools, teachers and pupils, together with descriptions of observed practices. These examples were collated and then foregrounded for exemplification based on the following criteria:

- consistent with principles of good assessment practice as discussed in the literature review;
- ~~as far as possible,~~ examples should have been ~~seen in action by tutors to ensure authenticity~~authenticated by science subject leaders or members of the research team;
- have visual clarity, but focus on the quality of content not presentation;
- samples of pupil work should be ~~genuine, complete with any errors~~originals rather than 'fair copies', demonstrating any alterations made following feedback;
- there should be more than one example for each statement and work should reflect a diversity of schools and contexts;
- the range of examples provided should support multimodal recording and creative practice;
- ethical processes have been enacted: schools are named as their work is celebrated, children are anonymous, parental permission has been received for images of children and the schools have approved all examples included prior to publication.

In order to fulfil the above criteria and provide broader exemplification, examples were sought beyond the immediate project schools, including from the 'expert' groups referred to above (n=43). All exemplification of authentic practices in real-school contexts (n=95) were hyperlinked to the relevant statements in a published, online version of the data-flow pyramid (Figure 5, see

pstt.org.uk/application/files/6314/5761/9877/taps-pyramid-final.pdf).

[insert figure 5 here]

Phase 4: Validation

Since DBR can include recursively nested research processes (Easterday et al., 2014), we elected to use Kane's argument-based approach to validation (1990, 2013) to test and verify the data-flow pyramid model as an effective self-evaluation tool for primary school use. Kane's approach involves two stages: a *formative stage* during which researchers construct an 'interpretive argument' for validity based mainly on existing evidence, and a *summative stage* during which the interpretive argument is subjected to empirical challenge - particularly its problematic assumptions (Kane 1990: 24-5). The evidence required for validation is thus the evidence needed to evaluate the claims being made (Kane 2013: 448). The steps within our interpretive argument are as follows:

1. The TAPS data-flow pyramid model (Figure 5) is a valid elaboration and operationalisation of the original Nuffield Foundation assessment framework (Figure 1); this is largely an assertion based on informal feedback from teachers and other researchers following conference presentations.
2. It has participant-confirmed validity as a credible model of the types of science assessment requiring to be undertaken at classroom, year-group and school levels, together with the ways in which formative assessment data can pass between these levels, serving summative purposes as it does so. Our evidence for this stage of the argument is drawn from participant teacher comments during cluster days and school visits, as above.
3. When used as a whole-school self-evaluation tool it can provide a valid picture of the strengths and weaknesses of science assessment practice across a primary school, increasing assessment literacy in users and enabling targeted development of specific aspects. Participant schools' use of the tool in its various versions to 'RAG rate' their own systems, add examples and construct action plans provided some evidence for this stage of the argument.

We acknowledge that the above argument does not take into account critical questions concerning the validity and reliability of formative assessment data passed between levels and the extent to which these could be compromised by accountability pressures on teachers to ensure their pupils reach 'expected' outcomes. Such an argument would require an examination of assessment practices in schools outside the original sample using the tool, which is beyond the scope of this article. We were, however able to~~To~~ test step 1 of the non-critical interpretive validity argument above~~, we convened~~ by convening a validation panel in April 2016, consisting of three constituencies:

- members of the original expert group which authored the Nuffield Foundation (2012) report (n=7);
- expert primary science teachers - members of the PSTC who had not been directly involved in the development of the model (n=4);
- acknowledged experts in the field of primary assessment (n=5).

We asked group 1 about the origins of their original 'data-flow pyramid' model and whether the TAPS version remained true to the principles of assessment outlined in their report. They explained that they had sought a 'big picture' to provide an overview of the use of science assessment data for different purposes within and beyond primary schools. As a former ecologist, one member had tried the analogy of a pyramid of biomass within an ecosystem and found that by using 'energy' to represent the flow of information and 'biomass' to represent the evidence collected the analogy 'seemed to work well' and was adopted by

consensus. They confirmed that the TAPS model (Figure 5) was entirely consistent with the principles of assessment in their report and that the examples in the online version helped to 'make those principles more real' for teachers, contributing to a shared understanding of assessment processes without which the formative to summative transition - which may occur at any point between layers 2 and 4 - would not be effective, resulting in schools resorting to additional testing to provide the quantitative data required in layers 4 and 5.

Group 2 were asked about how the operationalisation of the original Nuffield Foundation assessment framework represented by the TAPS model might support classroom teachers, responding that it 'has shown teachers what assessment looks like', helping them to recognise valuable formative evidence in activities they are already conducting as part of everyday classroom practice. They viewed the model as reducing teachers' confusion about assessment. The emphasis on both procedural ('working scientifically') and conceptual understanding was felt to be appropriate, with particular commendation of summaries of what 'pupils can do' (S1) for parents. They did however view the formative to summative transition within the model as requiring further focus.

Group 3 were asked to identify which layers within the model they considered the most important; to which they responded that layer 1 - which had been added to the original Nuffield model - was of particular value in emphasising pupils' roles in their own and peers' assessment. They were also asked whether any adjustments were needed. This prompted a querying of the position of reporting to parents, which the group felt was now 'too high' in the pyramid, giving the impression that it would be driven by quantitative data of a binary 'achieved/not achieved' nature, rather than qualitative assessment of a child's experiences and the impact on their self-esteem as young scientists. They felt that parents and carers needed to be involved at a much earlier stage to provide personal insights into pupils, and that this should be separate from high-level data reporting. Group 1 confirmed that this had been a feature of their original model, in which 'narrative' reporting would take place more than once per year. Overall, group 3 regarded the TAPS model as a very useful overview, showing *'what it looks like for teachers to gather evidence and moderate judgements'*, contributing greatly to the manageability of the process.

To test step 2 of the interpretive validity argument we took the model to four nationally-recognised schools recommended by our funder the Primary Science Teaching Trust (PSTT), asking them to comment on its credibility as a framework to analyse their practice (Earle, 2015). Whilst each school's approach to science assessment differed, all four recognised the TAPS model as representing the practice to which they aspired. For example, each school noted the importance of a 'shared understanding' of progression in science learning (box M3). They believed progression grids of inquiry skills used for planning, assessment and moderation would support staff and pupils to assess formatively and summatively. Another indicator of the credibility of the model is provided by the download statistics for the online version of the data-flow pyramid, (see link above), which had been downloaded 6032 times by the end of March 2017, suggesting strong interest in our work. The wider impact on schools of the school self-evaluation tool published on line is being independently evaluated.

To test step 3 of the interpretive validity argument above project schools completed an Impact Survey (November 2015, n = 9) which included re-evaluation of their RAG rating. Schools commented on specific areas where practice had changed, identifying 6 red boxes and 10 amber boxes which they felt had moved to green, together with 3 red boxes which had moved to amber. All 9 schools agreed or strongly agreed that the

use of the pyramid had improved their assessment literacy, increasing their understanding of teacher assessment, reliability, validity and the relationship between summative and formative. Six of the 9 agreed that their colleagues had a clearer shared understanding of what to look for in children's science, whilst three felt that it had actually improved the validity and reliability of assessment.

A further empirical test was provided by interviews with 6 participants from the original projects schools in June 2016. The analysis suggested an emerging understanding of the potential for use of evidence collected for formative purposes being summarised for monitoring, tracking and reporting:

> I guess it's when we use our tracking system; that then becomes your summative assessment. But you've done the work before that. There are lots of different objectives within one unit on plants or one thing on electricity. By looking at each objective as formative assessment, you can then see what their overall understanding of that particular unit is, if that makes sense (participant 1).
>
> 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 (participant 2).
>
> … what we're doing now is doing ongoing formative assessments throughout a unit of work, and at the end of each piece of unit of work… and we use those judgments at the end of each unit of work—and that's both the working scientifically and the conceptual knowledge—to inform an annual judgment about that child, which then goes towards a summative statement that is passed on to the next teacher and then used as a summative statement for the end of the key stage, which had been used at the end of key stage for tracking purposes (participant 3).
>
> We feel strongly that even summative assessment has to have a formative purpose. We want teachers to be asking 'how will the child completing this focus assessment task help us to improve teaching and learning for them and for future cohorts?' Making a judgement about the level that individual children are working at is a secondary outcome (participant 4).

The validation process continues with the data-flow pyramid model now being tested in new contexts beyond England and beyond the primary age group.

Discussion

Through following a design-based research (DBR) approach, teachers and researchers in the TAPS project have become a community of practice (Wenger, 1998), developing a shared understanding of the nature and purposes of assessment in science at a time of rapid, externally-imposed change (Brill and Twist 2013). Although teachers remained at the heart of the project, circularity of claims (Shavelson et al, 2003) has been circumvented by supplementing the community of practice to include those beyond the project partnership. We acknowledge that in the initial design phases there was not an equal relationship between participants as the research team took the lead; some of the teachers initially called the project 'the TAPS course' indicating that they felt they were receiving training rather than acting as co-researchers. As the project progressed a stronger collaborative culture emerged (Anderson & Shattuck, 2012) as researchers and teachers worked

together to make classroom observations, collect data, test prototypes, seek out and edit exemplification material. A combination of reported and observed data was collected; teachers reflected on their practice during 'cluster day' meetings and researchers visited the teachers in school to observe lessons and collect examples of teacher planning, pupil work and assessment records. In addition to this close partnership, the 'community' also included invited contributors, observers, 'expert' practitioners and some of the original Nuffield Foundation (2012) authors who acted as critical friends and validators for the project.

One of the prerequisites for developing effective teacher assessment in schools is that teachers have sufficient 'assessment literacy' (Klenowski, 2011) to make wise decisions about which data to collect to represent pupils' scientific learning, how to collect them, how to involve pupils in the process, how to make valid and reliable judgements on the evidence available and put it to effective use whilst recognising its limitations. This demands judicious use of professional wisdom and takes time to develop. During the three years of collaborative development of the 'data-flow pyramid model', we as a research team (both university researchers and school-based colleagues) have needed to grapple with the complex relationship between formative and summative purposes of assessment as they feature in the various levels of the model. As discussed above, we debated whether the transition from formative to summative occurred between levels 1 (ongoing formative) and 2 (monitoring) or between 2 and 3 (reporting). However, in one interpretation of the word 'summative', it could be argued that judgements are being made by teachers (and pupils) every time an assessment is made (Taras, 2007), representing a 'snapshot' of pupil attainment in a particular aspect of scientific learning at a particular point in time. This can occur in the base level of the 'pyramid' model when pupils are assessing their own work and ideas against known criteria, or in the layer above where teachers are gathering evidence of pupils' learning from a range of sources. In order to take the formative actions of 'giving feedback' or 'adapting the pace, challenge and content' of lessons, teachers need to have made a summative 'snapshot' judgement of the state of pupils' learning at a particular time. However, in the more usual use of the word 'summative' to indicate a summary of available evidence collected over a period of time, teachers might be making a more nuanced judgement of progress leading up to a pre-defined assessment point in order to achieve greater validity (Gardner et al, 2010). Given that this would involve consideration of 'old' evidence - dating from perhaps several months ago since when the pupil concerned might be expected to have progressed - it would not have the same status as a 'snapshot in time'. So by developing the 'data-flow pyramid model' during the TAPS project, researchers and participants have come to a deeper, more sophisticated understanding of the complex relationship between the formative and summative purposes of assessment.

Conclusion

In this article we have argued that a design-based research (DBR) approach to theory-building and school improvement has led to the development and exemplification of a promising approach to school self-evaluation in relation to assessment in science. The intertwining of a developing understanding of the processes of assessment alongside product development - although challenging and complex - has through an iterative process led to an evidenced-based set of recommendations for practice. The 'data-flow pyramid' model and self-evaluation tool has undergone a process of 'rapid prototyping' (Tripp and Bichelmeyer, 1990) and has been thoroughly tested in authentic contexts. A process of cross-checking each iteration involved teachers' testing 'in the field' and applying their context-specific professional understanding of primary science assessment, which researchers cross-referenced with theoretical perspectives. This

development and verification process has, we argue, led to a clearer articulation of the role of teacher assessment in primary science as theorised in the model proposed by the Nuffield Foundation (2012) authors. The resulting self-evaluation tool has the potential to enable schools to evaluate their assessment practices and empower teachers to make secure judgement about children's learning.

Education research that seeks to support pedagogical development needs to take account of sociocultural insights that, whilst meanings and interpretations vary considerably across uniquely-situated schools and teachers, there are useful generalisations to be made across contexts (Mercer and Littleton, 2007). The TAPS project design in which individual cases of schools working with a researcher came together within a framework of collaborative project days enabled both local, particular interpretations of theory, but also common general themes and shared understandings to emerge. The product of this DBR process - the data-flow pyramid - is thus both a flexible tool for creative school use and a theoretical model located within the national overarching culture of primary science education and its political constraints (Cobb, 2003, p. 9). The way the TAPS data-flow pyramid model encapsulates a range of activities as a system of assessment recognises that theory in education involves complexity, diversity and the dynamic interconnectedness of different aims and processes. It recognises that children and teachers operate within schools which are in turn subject to demands and influences beyond them. Arguably, a limitation of the model from a systems perspective is that is it bounded by the school; it does not include the government and scrutinizing authorities directly. However, it does acknowledge political use of summative assessment as a driving force and manages this by exemplifying ways to make summative assessment a dependable summary of children's learning in science. It recognises that validity, reliability and manageability are not absolutes, but can be supported and balanced in different ways with trade-offs for different stakeholders. Rooted in pragmatism, and recognising the national and international demand for comparative data, the TAPS data-flow pyramid is designed to make a positive impact on practices that many educators value: formative assessment, pupils engaging in scientific enquiry and autonomy for teachers within a supportive and rigorous framework. Further research is required to validate the extent to which formative assessment data passed between levels in schools using the tool are enhanced in quality.

Acknowledgements
The authors acknowledge the support of the Primary Science Teaching Trust, UK.

Notes on Contributors
Dan Davies is Professor of Science and Technology Education and Dean of the Cardiff School of Education at Cardiff Metropolitan University, UK
Sarah Earle is Senior Lecturer in Primary Science Education at Bath Spa University, UK
Kendra McMahon is Senior Lecturer in Primary Science Education and Co-director of the Centre for Research in Science and Technology Learning and Education (CRISTLE) at Bath Spa University, UK
Alan Howe is Head of Department of Education and Childhood Studies and Co-director of the Centre for Research in Science and Technology Learning and Education (CRISTLE) at Bath Spa University, UK
Christopher Collier is Senior Lecturer in Primary Science Education at Bath Spa University, UK

References

Anderson, T., & Shattuck, J. (2012) Design-based research: a decade of progress in education research? *Educational Researcher, 41(1),* 16-25.

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

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

Blank, A. (2008)Where has the third core subject gone? *Primary Science, 105*, 4–6. Retrieved 9 September 2016 from https://www.ase.org.uk/journals/primary-science/2008/11/105/1072/PSR105Nov-Dec2008p4.pdf.

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

Brill, F. & 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.

Brown, A. (1992) Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings, *The Journal of the Learning Sciences, 2(2),* 141-178.

Bryk, A., Gomez, L. & Grunow, A. (2010). *Getting Ideas Into Action: Building Networked Improvement Communities in Education*. Stanford, CA, Carnegie Foundation for the Advancement of Teaching. Retrieved 9 September 2016 from http://www.carnegiefoundation.org/spotlight/webinar-bryk-gomez-building-networked-improvement-communities-in-education

Cobb, P., Confrey, J, diSessa, A., Lehrer, R. & Schauble, L. (2003) The Role of Design in Educational Research, *Educational Researcher, 32(1),* 9-13.

Coll, R., France, B. & Taylor, I. (2005) The role of models and analogies in science education: implications from research. *International Journal of Science Education, 27(2),* 183-198.

Collins, A., Joseph, D. & Bielaczyc, K. (2004) Design Research: Theoretical and Methodological Issues, *The Journal of the Learning Sciences, 13(1),* 15-42.

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

Department for Education (DfE) (2014). *Assessment Principles.* Retrieved 9 September 2016 from: www.gov.uk/government/uploads/system/uploads/attachment_data/file/304602/Assessment_Principles.pdf

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

Earle, S. (2015) An exploration of whole school assessment systems. *Primary Science 136*, 20-22.

Easterday, M., Rees Lewis, D & Gerber, E. (2014) Design-Based Research Process: Problems, Phases, and Applications, *Learning and Becoming in Practice: The International Conference of the Learning Sciences (ICLS) 2014.* University of Colorado Boulder 23-27 June. Retrieved 30 March 2017 from https://egerber.mech.northwestern.edu/wp-content/uploads/2014/10/DesignResearch_Methodology_ICLS_2014.pdf

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

Gipps, C. (1994). *Beyond Testing: Towards a Theory of Educational Assessment.* London, Falmer.

Gipps, C. V. and Murphy, P. (1994). *A Fair Test? Assessment, Achievement and Equity.* Milton Keynes, Open University Press.

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

Harlen, W. & Qualter, A. (2014). *The Teaching of Science in Primary Schools.* Abingdon, Routledge.

Hartas, D. (Ed.) (2010). *Educational Research and Inquiry*. London, Continuum.

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

Kane, M. (1990) *An Argument-based Approach to Validation; ACT Research Report Series 90-13.* Iowa City, The American College Testing Program.

Kane, M. (2013) The Argument-Based Approach to Validation. *School Psychology Review*, *42(4)*, 448 – 457.

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

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

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

McGuigan, L. & Russell, T. (2015) Using multimodal strategies to challenge early years children's essentialist beliefs. *Journal of Emergent Science, 9*, 35-41.

McIntyre, N. (2015). *Increasing Achievement in Science Education: Learning Lessons from Finland & Estonia*. Retrieved 9 September 2016 from http://www.ase.org.uk/documents/eis-xtra-increasing-achievement/

Mercer, N. & Littleton, K. (2007) *Dialogue and the Development of Children's Thinking, a Sociocultural Approach*. London, Routledge.

Newton, P. (2009) The reliability of results from national curriculum testing in England. *Educational Research, 51(2),* 181–212.

Nuffield Foundation (2012). *Developing policy, principles and practice in primary school science assessment*. London, Nuffield Foundation.

Riddell, R. (2016). *Equity, trust and the self-improving schools system.* London, Trentham Books.

PISA (2012) Key findings. Retrieved 9 September 2016 from: http://www.oecd.org/pisa/keyfindings/pisa-2012-results.htm

Shavelson, R., Phillips, D., Towne, L. & Feuer, M. (2003) On the Science of Education Design Studies, *Educational Researcher, 32(1),* 25-28.

Short, J. (2014) *Principled use of Assessment in Primary Education in the 21st Century.* Retrieved 9 September 2016 from: http://inspir-ed.net/resources/

Stake, R. (1995) *Multiple Case Study Analysis*. New York, Guildford Press.

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

Stiggins, R. J. (1992) Two disciplines of educational assessment. *Measurement and Evaluation in Counselling and Development, 26,* 93-104.

Stobart, G. (2012) Validity in formative assessment. In Gardner, J. (ed) *Assessment and Learning, 2nd Edition*. London, Sage.

Taras, M. (2007) Assessment for Learning: understanding theory to improve practice, *Journal of Further and Higher Education, 31(4),* 363-371.

Taras, M. (2009) Summative assessment: the missing link for formative assessment, *Journal of Further and Higher Education, 33(1),* 57-69.

Task Group on Assessment and Testing (TGAT) (1988). *Task Group on Assessment and Testing: A Report.* London, DES.

Tripp, S. & Bichelmeyer, B. (1990) Rapid prototyping: An alternative instructional design strategy, *Educational Technology Research and Development, 38(1),* 31-44.

Wenger, E. (1998) Communities of practice: Learning as a social system. *Systems thinker, 9(5),* 2-3.

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

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

Wiliam, D. & Black, P. (1996) Meaning and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal* 22(5), 537-48.