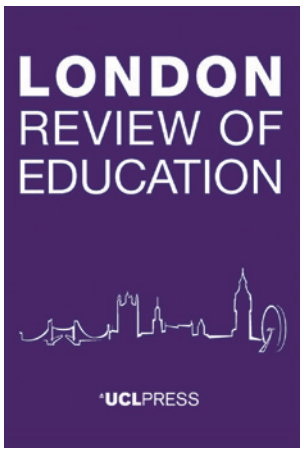

UCLPRESS

LONDON REVIEW OF EDUCATION

e-ISSN: 1474-8479

Journal homepage:
https://www.uclpress.co.uk/pages/london-review-of-education

Balancing the demands of validity and reliability in practice: Case study of a changing system of primary science summative assessment

Sarah Earle

How to cite this article
Earle, S. (2020) 'Balancing the demands of validity and reliability in practice: Case study of a changing system of primary science summative assessment'. *London Review of Education*, 18 (2): 221–235. https://doi.org/10.14324/LRE.18.2.06

Submission date: 19 August 2019
Acceptance date: 31 October 2019
Publication date: 21 July 2020

Peer review
This article has been peer-reviewed through the journal's standard double-blind peer review, where both the reviewers and authors are anonymized during review.

LONDON REVIEW OF EDUCATION

Earle, S. (2020) 'Balancing the demands of validity and reliability in practice: Case study of a changing system of primary science summative assessment'. *London Review of Education*, 18 (2): 221–235. DOI https://doi.org/10.14324/LRE.18.2.06

Balancing the demands of validity and reliability in practice: Case study of a changing system of primary science summative assessment

Sarah Earle* – *Bath Spa University, UK*

Abstract

Teacher summative judgements of children's attainment in science, which are statutory at age 11 in England, require consideration of both valid sampling of the construct and reliable comparison of outcomes. In order to develop understanding of the enacted 'trade off' between validity and reliability, this three-year case study, within the Teacher Assessment in Primary Science (TAPS) project, was undertaken during a period of statutory assessment change in England. The case demonstrates an ongoing balancing act between the demands of reliability and validity, and resulted in the development of a teacher assessment seesaw, which provides a model for both interpreting and supporting practice, within and beyond primary science.

Keywords: teacher assessment, validity, reliability, teacher assessment literacy, primary science

Introduction

This study examines the development of summative assessment practices in action and over time in a primary school in England, in order to explore how teachers attempt to balance the demands of validity and reliability in the context of policy-driven changes in practice. It is proposed that understanding of the way such a balancing act is conceived and enacted will enable future guidance for practice to be more closely tailored to the needs of teachers, enabling theories to do 'real work' in education (Cobb *et al.*, 2003). By supporting teachers with their assessment practices, pupils will benefit from more focused teaching that can 'address areas of misunderstanding or gaps in knowledge' (Harrison and Howard, 2009: 5). The term 'teacher assessment' is used to mean the assessment of pupils' learning by teachers, exemplified in primary science in this article, from which recommendations are relevant more broadly.

In line with a range of other countries (for example, Finland, Australia, New Zealand, Scotland and Wales), the summative assessment of primary science in England relies on teacher assessment (since the removal of national testing in 2009). However, the past 10 years have seen significant changes in the structure of assessment for English schools, moving from a system of 'levels' (DfEE and QCA, 1999; DES and the Welsh Office, 1988) to Age Related Expectations (DfE, 2013a, implemented 2014/15). The long-standing levels system required teachers to match each child's attainment to a broad level descriptor, while the Age Related Expectations contained a list of detailed criteria that needed to be 'met'. Schools were encouraged to develop their

*Email: s.earle@bathspa.ac.uk

own approaches to summative assessment under this new system (DfE, 2013b). The case study period (2013–16) was selected in order to closely examine assessment practices at a time of statutory change, and to consider the way the school managed the demands of validity and reliability. The case study is also set within the Teacher Assessment in Primary Science (TAPS) project, which employs a design-based research approach using collaborative and iterative cycles for development of support for practice (Davies *et al.*, 2017).

Validity and reliability in teacher assessment

Validity concerns whether an assessment is actually assessing what it claims to be, and the extent to which it is fit for purpose (Green and Oates, 2009). It 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). For example, when considering the validity of a summative assessment of primary science, a key question is whether it effectively samples enough of the domain to be representative (ibid.). Primary science is about making sense of the world. To do this, the individual needs appropriate attitudes, skills, knowledge and understanding; thus, science is both a body of knowledge and a process of inquiry. Inquiry skills are 'not well defined constructs' (Millar, 2010: 127) and are also embedded within the context in which they are applied, making judgements of both validity and reliability in primary science assessment problematic.

Reliability concerns trust in the accuracy or consistency of an assessment (Mansell *et al.*, 2009), for example, whether the same judgement would be made if the task was given on a different occasion or marked by a different teacher (Newton, 2009). In primary science, this poses a range of difficulties, with young children's expressions of ideas affected by the mode of assessment (for example, requiring written or oral answers), and the use of skills being heavily dependent on the context (for example, drawing conclusions from different inquiries). Filer and Pollard (2000) caution that since school summative 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. 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, a summative assessment also needs to sample enough of the domain.

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. A pyramid-shaped 'formative to summative' model of teacher assessment was proposed by the Nuffield Foundation (2012), whereby the rich formative information gathered in the classroom could be used to make summative judgements at a later date. The TAPS project operationalized the pyramid model into a school self-evaluation tool (Davies *et al.*, 2017), detailing how a wide range of formative assessment strategies can be used to actively involve pupils and inform summative judgements. Harlen (2009), 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 under-representation of elements of the curriculum (construct under-representation), providing a broader

sampling of the construct by taking into account the wide range of information available in the classroom.

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) because of the lack of opportunities to compare judgements. Johnson (2013) also questions reliability in teacher assessment, noting the limited and ambiguous research in this area. One concern is the permanence of assessment evidence, since recorded outcomes can be considered for consistency by other 'raters', but outcomes from activities such as group discussion are harder to capture and compare. In their seminal paper, Wiliam and Black (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. Connolly *et al*. (2012), with Queensland's long history of moderation, 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, 2015a, 2015b), where teachers make judgements based on a wide range of evidence, could be open to concerns regarding subjectivity and teacher bias (Campbell, 2015).

Wiliam (2003) argues that there is inevitably a 'trade off' between reliability and validity when assessing summatively. Halliday (2010: 370) suggests that a 'trade off' between reliability and validity is necessary, since reliability relies on a narrowing of task variables to support agreement of those marking the task, while validity depends on the opposite: as broad a sampling of the subject as possible. Sadler (1989: 122) asserts that validity should take precedence when the aim is formative, for diagnosis and improvement. 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.

With 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). 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 shared understanding of the domain. Klenowski and Wyatt-Smith (2014) explain 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). Teachers could learn to use more holistic judgements rather than rely on a prescriptive tick list (ibid.: 458). However, 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). In addition, the large-scale collection of evidence to support such teacher assessment, prompts questions of manageability for teachers.

This article draws upon a discrete three-year case study, which formed part of the ongoing TAPS project, to answer the following research questions:

(1) How does a school system of primary science summative assessment address the validity/reliability trade-off over time?
(2) How can study of changes over time in summative teacher assessment be used to inform guidance for practice?

Methods

This empirical enquiry is placed within interpretative and applied research traditions, engaging with a real-world setting to develop both theory and practice, utilizing a design-based research (DBR) methodology to engineer products and recommendations to inform practice (Brown, 1992: 143). DBR involves collaborative partnership between researchers and practitioners (Anderson and Shattuck, 2012) during iterative cycles or phases. In this study, the university tutors structured project development days to provide opportunities for discussion of current school practices and policy, alongside consideration of assessment principles (Nuffield Foundation, 2012). During this period of statutory changes to assessment, it had become unclear as to what constituted 'best practice', provoking a re-evaluation of current practice. Practitioners and researchers analysed whether practice could be aligned with the proposals in the Nuffield 'formative to summative' model, operationalizing the model into a self-evaluation tool (Davies *et al.*, 2017). It should be noted, therefore, that changing practices in the case study school were intertwined and part of the research project. This study examines how school practice changed within this context of collaborative research.

Collins *et al*. (2004) note the importance of multiple ways of looking, in order to consider the many layers of the school learning environment, which in this study was accomplished by a three-year case study. The case study is a 'study of an instance in action' (Adelman *et al.*, 1976: 141), and it is understanding of the 'in action' element that is so central to DBR and to this study of assessment, where the practice is not 'frozen' (Cohen *et al.*, 2011).

The case was selected to be informative rather than representative (ibid.). This purposive sampling was driven by the research questions, which required exploration of change over time; the goal was depth rather than breadth (Mears, 2012). Participation in an in-depth study over three years already suggests that the school is atypical; such participation requires the support of the head teacher and the subject leader 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, but to answer the research questions, the 'right source' was needed (Newby, 2010), a school that could commit to long-term involvement.

School B was selected from the TAPS project group because it provided the most complete case record for changes over time to be explored, being one of the few schools that did not have a change of head teacher or subject leader during the project. School B is a one-form entry primary school in England where attainment at age 11 was reported as higher than the national average. 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 (an indicator of low socio-economic status) was below the national average.

The data for School B were collected between March 2013 and June 2016 using the range of methods described in Table 1. The iterative DBR cycles alternated

Table 1: Overview of data collection methods

School B data collection methods	Collection points March 2013–June 2016	Number of items in case record
Documentation: e.g. Primary Science Quality Mark submission*, documents collected on visits: policies, lesson plans, records, work samples	6 school visits 2 PSQM submissions	58
Non-participant observation: e.g. Lesson observation using both a project observation schedule and field notes, observation of meeting/presentation using field notes	3 lessons 1 staff meeting 2 presentations	11
Semi-structured (researcher-led) discussions or meetings: e.g. interview/meeting, group discussion	3 interviews 4 group discussions	4
Written tasks (researcher-led): e.g. completion of questionnaire, sorting activity, self-evaluation on project development days.	8 development days	13
Total items in case record		**86**

*The Primary Science Quality Mark is an award scheme requiring evidence to be uploaded after a year of school development (www.psqm.org.uk)

between school visits and project development days. The school visits included science lesson observations and semi-structured interviews with school staff. The project development days provided the opportunity for all of the project schools and tutors to discuss practice, findings, draft resources and guidance. Two or three members of staff attended development days from each school, including the science subject leader, assessment coordinator and head teacher. Nevertheless, the majority of the data comes from the subject leader as the 'gatekeeper' for science in the school, since they lead school policy and staff training, and provide support and guidance to colleagues. She was also responsible for writing the school's evidence submission for the Primary Science Quality Mark (PSQM). 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.
- Time triangulation: ongoing involvement and data collection with each school for a three-year period.
- Investigator triangulation: on two occasions different researchers collected data (first school visit and final subject leader interview).
- Source triangulation: involving a number of teachers from the school (although the majority of the data came from the subject leader).

In addition, respondent validation was employed by sharing the data, findings and interpretations with participants. The study was also placed within a larger research

project, providing prolonged engagement (Lincoln and Guba, 1985) and the opportunity for methods, findings and interpretations to be discussed and 'tested' with the wider research team throughout the process.

Ethical principles of voluntary informed choice, consent and 'right to withdraw' were followed (BERA, 2011), with ongoing ethical discussion and decision making with participants taking place throughout the research process (Luttrell, 2010). The case study data were anonymized and stored securely.

A qualitative content analysis approach was taken (Silverman, 2011) supported by ATLAS.ti software for coding and retrieval of data. The coding of the data began with a list of theory-led codes generated from the research questions and literature (Nuffield Foundation, 2012; Davies *et al.*, 2014); further emergent codes were also added as they arose from the data. The 'code and retrieve' use of the software enabled 'to and fro' between the raw data and interpretations of them, with efficient 'constant comparison' (Robson, 2011) providing rigour to the analysis. Codes and items were revisited a number of times to ensure consistency across the data set. 'Higher order codes' or themes emerged from the case study data, which were both recurring and pertinent to the research questions (for further detail, including code frequencies, see Earle, 2018). In the discussion below, the 'higher order' **codes** are written in bold on their first occurrence in a section, to support transparency of data analysis.

In order to compare changes over time, the case record was organized into the three DBR phases, as detailed in Table 2. The DBR phase structure allowed for comparison between the frequency of codes at each phase, which could represent shifts in focus for the subject leader or school, together with checks to avoid overemphasis on the 'loudest or brightest' data (Cohen *et al.*, 2011).

Table 2: Design-based research phases for analysis

Phase	Dates	Data identifier
1: Exploration	March 13–November 13	B1 to B21
2: Development	February 14–January 15	B22 to B53
3: Implementation	March 15–June 16	B54 to B86

Results

Phase 1: Searching for consistency

In Phase 1, the research focused on exploration of current practices, finding that teacher concerns relating to reliability were evident, with a focus on consistency, criteria structures, levelling (assigning a level based on national criteria), paper evidence and marking. For example, the subject leader focuses on criteria **structures** and **evidence** when writing an explanation of assessment practices:

> Extract 1
>
> To try to 'standardise' summative assessments, some published material is used, including past test papers. Our teachers are using a range of materials to inform judgements including: Scheme of Work A, Scheme of Work B, government guidance A, local government guidance B, Tests A, Tests B – and other materials found on internet … (Subject leader written description, June 2013, B3)

The subject leader describes how multiple published materials or **structures** were used to try to *'standardise'* their summative assessments, indicating a concern for reliability. The long list of supportive structures raises questions of manageability for staff if they are required to use all of these resources, together with possible issues with reliability if there are differences between the criteria for each. In Phase 1, summative assessment appeared almost synonymous with **levelling**, with staff development focused on gaining confidence with levelling (PSQM reflection, March 2013, B1).

Consistency was noted by the subject leader as a key issue, and in an interview the subject leader commented that: 'core principles for assessment in science have been established but there is difference in practice amongst classes' (November 2013, B10). Concerns regarding 'consistency' as an issue was predominantly coded in Phase 1 (12 out of 15 occurrences), suggesting that it became less of a concern later in the case record.

A range of **strategies** to elicit and record pupil ideas were represented in the case record. One recurring theme in Phase 1 was a teacher focus on **marking** and children responding to marking. The school's marking policy at the time 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 (November 2013, B14–15) and on subsequent school visits (February 2014, B22). The detailed marking raises questions about manageability (Independent Teacher Workload Review Group, 2016), together with the value placed on written recording, which may be a particular issue for younger children.

Phase 2: Evidence and moderation

During Phase 2, the concern for **evidence** collection and recording continued to be a school focus (school visits B22). Moderation is a key way to improve reliability in teacher assessment (Harlen, 2007), as discussed at project development days, and School B's **moderation** staff meeting included providing dedicated time for the teachers to discuss how they were making their judgements (see Extract 2). It was not a simple checking of levels assigned to individual pieces of work; the aim appeared more to be to develop the assessment literacy of the staff, to make explicit the tacit knowledge of how to make judgements (Sharpe, 2004):

Extract 2

What's required to level a piece of work?

1. Useful to see planning and know the context.
2. Need to know what type of support might have affected outcomes.
3. Useful to capture verbal comments and observe contributions in sessions.
4. Clear assessment criteria. Agreed sources.
5. Good knowledge of curriculum content and level descriptors.
6. 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 handout, June 2014, B43)

During the staff meeting, the teachers were exploring what they needed to be able to 'level a child', that is, to ascribe a summative grade. The suggestion of utilizing: 'a greater range of examples' and looking for ways to capture oral pupil talk, could enhance validity, with the summative judgement based on information that 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) 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 the Nuffield Foundation (2012) and the TAPS pyramid (Davies *et al.*, 2014) recommendations, which had been discussed at development days, that information gathered for formative purposes could be summarized 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 summarizing formative assessment and repeated summative assessment.

Phase 2: Making assessment manageable

School B explored a range of strategies over the case study period, which had been collated by project schools during development days, in the attempt to find workable, manageable solutions. One of these was to narrow the focus for teacher attention by predefining success criteria (B29) and expectations for pupil outcomes in the form of **differentiation** (B30). For example, in a Year 5/6 (age 9–11) lesson on mixing materials, pupils worked in predesignated attainment groups and with group-specific recording sheets. The recording sheets provided a different amount of structuring and challenge for different groups, with the simplest sheet directing the children to identify if a new product was made, while the more complex sheet asked the children to explain the changes to the materials (B32, B33). Such structuring or scaffolding of the level of challenge within the lesson is one of the features of Assessment for Learning (Loughland and Kilpatrick, 2015). However, by matching the level of challenge to the groups before the lesson, the teacher had pre-decided the pupil outcomes; for example, the pre-prepared end of lesson expectation grid already had pupil names typed underneath, ready to be ticked (levelled planning, B29). The information gathered in the lesson 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 pre-existing summative judgements.

Phase 3: Range of information

In Phase 3, the subject leader describes the supportive nature of moderation and 'sharing practice sessions':

> Extract 3
>
> 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. (PSQM C2 reflection, March 2015, B55)

There still appears to be a tension with regard to consistency – how much to stick to the *'framework'* and how much *'freedom for teachers to try what works'*. However, the

'*sharing and discussing with each other about what is being tried*', which had been a feature of the development days, 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 1, B3). By 2016, the subject leader advocates a different approach: summative assessment that was not described separately, or based on separate materials, but was ongoing and informed by formative assessment:

> Extract 4
>
> 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. (Subject leader presentation planning, May 2016, B80)

The subject leader 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.

Phase 3: Confidence in teacher judgement

In Phases 1 and 2, the subject leader listed a range of **structures** to support summative assessment, and there was an emphasis on **records** and **evidence**. In Phase 3, there was more focus on the role of the teacher:

> Extract 5
>
> 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. (Development day 7, November 2015, B78)

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 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 subject leader 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 point was also noted in a subject leader interview:

> Extract 6
>
> 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. We've found more efficient ways of doing that … 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 hasn't become onerous; it has become upskilled, I'd say. (Subject leader interview, June 2016, B83)

The subject leader suggested that their collection of evidence became more manageable; for example, rather than writing down all pupil comments, 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. While 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, as noted in Phase 2 above.

Phase 3: More open

There is some evidence that differentiation became less closed, with children choosing their level of challenge for homework activities (Spring 2015, B66), and that grouping became more mixed rather than by prior attainment (May 2016, B81). The subject leader also described how pupils are given more *'opportunities'* to demonstrate their understanding and independence, particularly via *'open-ended inquiry'* (June 2016, B83). The opening out of activities could enhance validity, providing for a wider range of pupil outcomes and more divergent assessment (Torrance and Prior, 1998).

Discussion

In answer to the first research question, regarding the development of summative assessment practice over time, the data suggest that School B placed emphasis on different elements of reliability and validity over time. This discussion will utilize a teacher assessment seesaw model (adapted from Earle, 2017) to support analysis of the ongoing balancing of validity and reliability seen in the case study, providing a model that can be applied more broadly (Research question 2).

Initially, in Phase 1, there was a focus on reliability in terms of collection of written evidence and checking against multiple criteria structures. The ongoing concern for consistency and evidence appeared to prioritize reliability, perhaps at the cost of validity. This also had low manageability because of the number of different structures used for cross-checking, as represented in Figure 1. There appeared to be less of a focus on validity, with consideration of only written paper formats, which may provide a limited sampling of practical primary science, together with being problematic for the younger children in the school.

The ongoing focus on evidence arguably led to repeated summative judgements, a concern to 'level' or judge at each interaction. Alternatively, the emphasis on

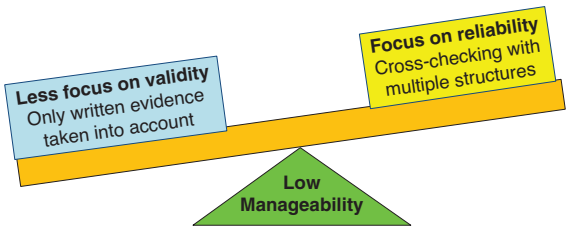

Figure 1: Phase 1 Summative teacher assessment seesaw: Focus on reliability

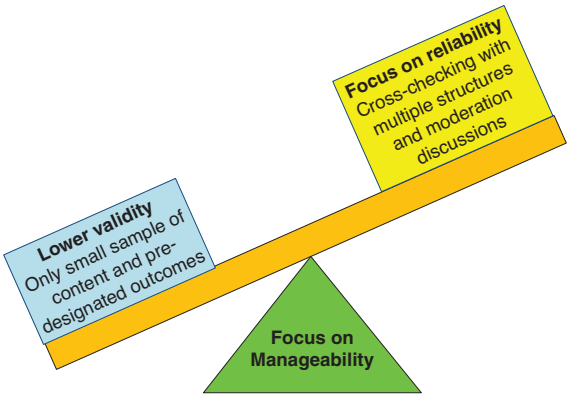

Figure 2: Phase 2 summative teacher assessment seesaw: Focus on reliability and manageability

evidence 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: 176) found improvements at the introduction of statutory summative teacher assessment, where practices moved from an intuitive approach to one based on evidence and written records.

During Phase 2, the concern for reliability continued, with the teachers taking part in moderation discussions and trialling strategies to develop their understanding of assessment within the subject (as represented in Figure 2), but an additional focus on manageability led to predetermined pupil outcomes via grouping or differentiated recording. Some emphasis on evidence to support reliable judgements is necessary, but an overemphasis on evidence may have a negative effect on manageability and validity, if assessments become too closed.

In Phase 3, considerations of teacher assessment literacy and broadening the range of outcomes signified a focus on validity. The data indicate more open tasks, without pre- assigned outcomes, and a wider range of information utilized to inform assessments, recognizing that written tasks provide only one form of evidence in practical primary science. Summative assessment came to be conceptualized as less of a 'bolt on' (B76) and more of an attainment summary, informed by a range of information. The increased confidence in teacher assessment described by the subject leader is represented in Figure 3 by an understanding and balancing of the demands of validity and reliability, recognizing that both need to be considered in developments of assessment practice.

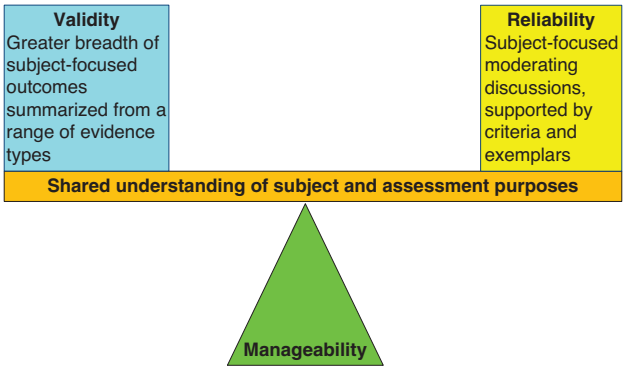

Figure 3: Phase 3 summative teacher assessment seesaw: Focus on validity

Phase 3 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 reconceptualized as a developmental process. When the subject leader commented that 'hearing a child is valid' for example (B78), it suggested development in teacher assessment literacy: a broadening in understanding of the types of information that can be used for assessment, which could lead to a wider sampling of the curriculum. Nevertheless, 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, for without an understanding of progression within the subject then it would be difficult to make a judgement or decide a next step.

The teacher assessment seesaw model in Figures 1 to 3 provides a very simplified view of practice in School B, but it is also proposed that such a simplification could support teacher assessment literacy by active engagement in discussion of key concepts (DeLuca and Johnson, 2017). In line with DBR principles, this model aims to do more than describe the problem for teacher assessment, the aim is to provide a tool to support teacher understanding and practice in application of assessment principles such as validity and reliability. In Figure 3, the model depicts:

- **Validity** focuses on content validity, with the aim of providing a summary of the child's performance throughout the whole of the curriculum, to combat construct under-representation. By basing summative reporting on a range of evidence types, construct irrelevance, related to specific ways of collecting pupil outcomes, could also be reduced (Black and Wiliam, 2012).
- **Reliability** requires reference to criteria, exemplars and moderating discussions, which support consistency and confidence in judgements. The discussions should be focused on the curricular objectives to avoid unconscious bias from assumptions about the child's behaviour or performance (Campbell, 2015).
- **Manageability** is a key component for teachers, because if implementation is too demanding then the assessment system 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.

Figure 3 provides a stimulus for professional dialogue around the validity–reliability 'trade off', recognizing that it is not possible to have a highly repeatable, standardized assessment that samples the whole of practical primary science. It should be noted that School B took three years to reach more of a 'balance' in summative assessment practice. Teachers needed 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. In addition, it should be recognized that this is an ongoing balancing act for schools, within a changing education system. Changes in staffing, curriculum guidance, statutory assessment procedures and inspectorate focus areas are all likely to affect the balance of assessment practice.

Harlen (2007) asserts that summative teacher assessment can be as reliable as it needs to be with moderation. Although the meaning of 'moderation' may need to be clarified, with some referring to a process whereby judgements were checked for inter-rater reliability (Johnson, 2013), and others referring to a process of professional dialogue whereby the meaning of criteria or types of evidence were explored, as seen in School B's staff meeting. Connolly *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. A shared understanding across the school, of science and of assessment, appeared to be enhanced by a criterion structure and moderation discussions. This shared understanding or shared criteria meant that formative assessment could be summarized for summative purposes because both assessments were using the same benchmarks for decision making.

In order to develop teacher assessment literacy, there is a need to recognize 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, particularly in a policy-driven system where concerns for reliability dominate. The teacher assessment seesaw model is presented as a way of representing the balancing act between validity concerns, which advocate basing judgements on a broad range of information, and reliability concerns, which require shared criteria, exemplars and moderation.

Acknowledgements

I thank the Primary Science Teaching Trust for funding the Teacher Assessment in Primary Science (TAPS) project and the Primary Science Quality Mark for supporting access to their database of award submissions.

Note on the contributor

Sarah Earle has led the Teacher Assessment in Primary Science (TAPS) project at Bath Spa University since 2015, working in collaboration with schools across the UK. Prior to this, during her 13 years teaching in primary schools, she was an assessment coordinator and science subject leader, before moving to initial teacher education as a senior lecturer for primary PGCE at Bath Spa University in 2012.

References

Adelman, C., Jenkins, D. and Kemmis, S. (1976) 'Re-thinking case study: Notes from the second Cambridge Conference'. *Cambridge Journal of Education*, 6 (3), 139–50. https://doi.org/10.1080/0305764760060306.

Anderson, T. and Shattuck, J. (2012) 'Design-based research: A decade of progress in education research?'. *Educational Researcher*, 41 (1), 16–25. https://doi.org/10.3102%2F0013189X11428813.

BERA (British Educational Research Association) (2011) *Ethical Guidelines for Educational Research*. Rev. ed. London: British Educational Research Association.

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), 451–69. https://doi.org/10.1080/0969594X.2011.557020.

Black, P. and Wiliam, D. (2012) 'The reliability of assessments'. In Gardner, J. (ed.) *Assessment and Learning*. 2nd ed. London: SAGE Publications, 243–63.

Brown, A.L. (1992) 'Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings'. *Journal of the Learning Sciences*, 2 (2), 141–78. https://doi.org/10.1207/s15327809jls0202_2.

Campbell, T. (2015) 'Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment'. *Journal of Social Policy*, 44 (3), 517–47. https://doi.org/10.1017/S0047279415000227.

Cobb, P., Confrey, J, DiSessa, A., Lehrer, R. and Schauble, L. (2003) 'Design experiments in educational research'. *Educational Researcher*, 32 (1), 9–13. https://doi.org/10.3102%2F0013189X032001009.

Cohen, L., Manion, L. and Morrison, K. (2011) *Research Methods in Education*. 7th ed. London: Routledge.

Collins, A., Joseph, D. and Bielaczyc, K. (2004) 'Design research: Theoretical and methodological issues'. *Journal of the Learning Sciences*, 13 (1), 15–42. https://doi.org/10.1207/s15327809jls1301_2.

Connolly, S., Klenowski, V. and Wyatt-Smith, C.M. (2012) 'Moderation and consistency of teacher judgement: Teachers' views'. *British Educational Research Journal*, 38 (4), 593–614. https://doi.org/10.1080/01411926.2011.569006.

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.J., 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), 1869–90. https://doi.org/10.1080/09500693.2017.1356942.

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 and Practice*, 24 (2), 121–6. https://doi.org/10.1080/0969594X.2017.1297010.

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), 248–66. https://doi.org/10.1080/10627197.2016.1236677.

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

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

DfE (Department for Education) (2013b) 'Assessing without levels'. Online. https://tinyurl.com/ttv2h82 (accessed 24 November 2019).

DfEE (Department for Education and Employment) and QCA (Qualifications and Curriculum Authority) (1999) *The National Curriculum: Handbook for primary teachers in England*. London: Department for Education and Employment.

Earle, S. (2017) 'The challenge of balancing key principles in teacher assessment'. *Journal of Emergent Science*, 12, 41–7.

Earle, S. (2018) 'The Relationship between Formative and Summative Teacher Assessment of Primary Science in England'. Unpublished PhD thesis, Bath Spa University.

Filer, A. and Pollard, A. (2000) *The Social World of Pupil Assessment: Process and contexts of primary schooling*. London: Continuum.

Gardner, J., Harlen, W., Hayward, L., Stobart, G. and Montgomery, M. (2010) *Developing Teacher Assessment*. Maidenhead: Open University Press.

Gipps, C., Brown, M., McCallum, B. and McAlister, S. (1995) *Intuition or Evidence? Teachers and National Assessment of Seven Year Olds*. Buckingham: Open University Press.

Green, S. and Oates, T. (2009) 'Considering alternatives to national assessment arrangements in England: Possibilities and opportunities'. *Educational Research*, 51 (2), 229–45. https://doi.org/10.1080/00131880902891503.

Halliday, J. (2010) 'Educational assessment'. In Bailey, R., Barrow, R., Carr, D. and McCarthy, C. (eds) *The Sage Handbook of Philosophy of Education*, 369–84. London: SAGE Publications.

Harlen, W. (2007) *Assessment of Learning*. London: SAGE Publications.

Harlen, W. (2009) 'Improving assessment *of* learning and *for* learning'. *Education 3–13: International Journal of Primary, Elementary and Early Years Education*, 37 (3), 247–57. https://doi.org/10.1080/03004270802442334.

Harrison, C. and Howard, S. (2009) *Inside the Primary Black Box: Assessment for learning in primary and early years classrooms*. London: GL Assessment.

Independent Teacher Workload Review Group (2016) *Eliminating Unnecessary Workload around Marking: Report of the Independent Teacher Workload Review Group*. London: Department for Education.

Johnson, S. (2012) *Assessing Learning in the Primary Classroom*. London: Routledge. Johnson, S. (2013) 'On the reliability of high-stakes teacher assessment'. *Research Papers in Education*, 28 (1), 91–105. https://doi.org/10.1080/02671522.2012.754229.

Klenowski, V. and Wyatt-Smith, C. (2014) *Assessment for Education: Standards, judgement and moderation*. London: SAGE Publications.

Lincoln, Y.S. and Guba, E.G. (1985) *Naturalistic Inquiry*. Newbury Park, CA: SAGE Publications.

Loughland, T. and Kilpatrick, L. (2015) 'Formative assessment in primary science'. *Education 3–13: International Journal of Primary, Elementary and Early Years Education*, 43 (2), 128–41. https://doi.org/10.1080/03004279.2013.767850.

Lum, G. (2015a) 'Introduction'. In Davis, A. and Winch, C. *Educational Assessment on Trial*. Ed. Lum, G. London: Bloomsbury Academic, 1–6.

Lum, G. (2015b) 'Afterword: Can the two positions be reconciled?'. In Davis, A. and Winch, C. *Educational Assessment on Trial*. Ed. Lum, G. London: Bloomsbury Academic, 107–34.

Luttrell, W. (ed.) (2010) *Qualitative Educational Research: Readings in reflexive methodology and transformative practice*. New York: Routledge.

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

Mears, C.L. (2012) 'In-depth interviews'. In Arthur, J., Waring, M., Coe, R. and Hedges, L. (eds) *Research Methods and Methodologies in Education*. London: SAGE Publications, 170–6.

Millar, R. (2010) 'Practical work'. In Osborne, J. and Dillon, J. (eds) *Good Practice in Science Teaching: What research has to say*. 2nd ed. Maidenhead: Open University Press, 108–34.

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

Newton, P.E. (2009) 'The reliability of results from national curriculum testing in England'. *Educational Research*, 51 (2), 181–212. https://doi.org/10.1080/00131880902891404.

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

Robson, C. (2011) *Real World Research*. 3rd ed. Chichester: Wiley.

Sadler, D.R. (1989) 'Formative assessment and the design of instructional systems'. *Instructional Science*, 18 (2), 119–44.

Sharpe, R. (2004) 'How do professionals learn and develop? Implications for staff and educational developers'. In Baume, D. and Kahn, P. (eds) *Enhancing Staff and Educational Development*. London: RoutledgeFalmer, 132–53.

Silverman, D. (2011) *Interpreting Qualitative Data: A guide to principles of qualitative research*. 4th ed. London: SAGE Publications.

Stobart, G. (2009) 'Determining validity in national curriculum assessments'. *Educational Research*, 51 (2), 161–79. https://doi.org/10.1080/00131880902891305.

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 (4), 221–58. https://doi.org/10.1007/s10833-006-0001-8.

Taras, M. (2005) 'Assessment – summative and formative – some theoretical reflections'. *British Journal of Educational Studies*, 53 (4), 466–78. https://doi.org/10.1111/j.1467-8527.2005.00307.x.

Torrance, H. and Pryor, J. (1998) *Investigating Formative Assessment: Teaching, learning and assessment in the classroom*. Buckingham: Open University Press.

Wiliam, D. (2003) 'National curriculum assessment: How to make it better'. *Research Papers in Education*, 18 (2), 129–36. https://doi.org/10.1080/0267152032000081896.

Wiliam, D. and Black, P. (1996) 'Meanings and consequences: A basis for distinguishing formative and summative functions of assessment?'. *British Educational Research Journal*, 22 (5), 537–48. https://doi.org/10.1080/0141192960220502.