Teach First

Assessment

Research review 2022-23

Contents

Introduction

A lot of school improvement focus is on results, particularly SATs and GCSEs. Secondary schools are ranked on the basis of attainment and progress 8 measures, primary schools on progress measures that are a type of 'value added' measure. As a consequence, many school leaders spend a lot of time collecting and analysing data in a bid to identify areas of weakness and direct strategic improvement. The difficulty is that the validity of the inferences that can be made from such summative data, is limited.

"Look, we would all love our numbers to be smooth lines up into the light... but that's not how it works ...

Those numbers are output measures ... I think that if you focus on the inputs... in the long term you get better results." (Business Insider, 2014)

Essentially, assessment has been side-tracked by external demands, and its true purpose lost in spreadsheets and accountability measures.

This research review explores the available research around the purpose and design of assessment.

The purpose of assessment

Assessment needs to be reclaimed as a tool to be used by teachers in the classroom, as an aspect of subject-specific pedagogy. Rather than just a simple measurement tool, it can also be a means of growth and progress. Assessment helps determine what is considered valuable in terms of subject/topic choices, and therefore influences the decisions made about time allocation in the curriculum. Historical decisions around assessment have driven changes in teaching and learning, and vice versa. The centrality of GCSEs and the abolition of KS3 SATs can be seen in continued debates over the apportioning of time between KS3 and KS4. The alignment of assessment and teaching and learning, however, should not be taken for granted (James, 2006).

Assessments are used to generate information on which to make decisions about student learning. However, too often information collected for one purpose is also used to measure something entirely different. For example, assessments that only test a certain number of components of a subject may be used to measure overall attainment in that subject (Kime, 2017a). Assessment shapes the culture of schools as well as determining what is taught and how, influencing how both teachers and pupils think about learning (Earle, 2021).

Evidence Based Education outlines the four pillars of assessment as:

- purpose
- value
- validity
- reliability

and suggests three steps for robust assessment:

1. The construct: what is the specific skill, knowledge or understanding that we intend to assess?
2. The end use: what do we want to do with the information generated?
3. The best tool: what and when is the most appropriate, effective and efficient way to assess? (Kime, 2017a)

The effective use of assessment provides both the means to identify whether students have succeeded and the information to help teachers to support those who have not yet 'got there'.

(Earle, 2021)

Newton lists 18 categories of purpose for educational assessment judgements:

1. social evaluation uses
2. formative uses
3. student monitoring uses
4. transfer uses
5. placement uses
6. diagnosis uses
7. guidance uses
8. qualification uses
9. selection uses
10. licensing uses
11. school choice uses
12. institution monitoring uses
13. resource allocation uses
14. organisational intervention uses
15. programme evaluation uses
16. system monitoring uses
17. comparability uses
18. national accounting uses

He also indicates the possibility of a longer list as well as identifying different sub-purposes (Newton, 2007, pp.161-162).

Validity and reliability

There is no such thing as a 'valid assessment', only one that is valid for a specific purpose. We are referring to the ability of the assessment to test what it intends to measure, and to provide information which is both valuable and appropriate for the intended purpose. Validity is about the inferences you make based on the information generated (Kime, 2017c).

When we talk of validity and great assessments, we are referring to the assessment's ability to

support the claims we want to make based on the information generated.

(Kime, 2017c)

Reliability in the assessment of student learning relates to accuracy and consistency over time, and context. The two most important factors in the reliability of assessments are the precision of the tasks and questions used to prompt student responses, and the accuracy and consistency of the interpretations derived from the responses. Fundamentally, assessment is a proxy for something that can't be seen, meaning no assessment is 100% reliable or valid. There is always 'noise' (Kime, 2017b).

Any assessment involves a balance between validity and reliability, and the choices we make in the classroom are both determined and affected by this balance. With all students sitting the same test under the same conditions we improve reliability, but if we look to assess what the students already know we might increase validity. Converting outcomes into numerical data that can be entered into a tracker may define progress in percentages, but these may reflect coverage rather than representing a valid assessment of an individual's attainment (Earle, 2017)

Value

Given the challenges and workload associated with assessment, it carries a high opportunity cost. The value of investing time and effort in carrying out assessment must be reflected in how well it fulfils its intended purpose. Assessment can have both positive and negative effects: pupils studying more or high-quality feedback are positive 'washback', but unintended negatives can include workload increases, teaching to the test and decreased time for other activities (Kime, 2017d).

Moderation

Moderation is intended to develop a shared understanding around assessment decisions related to student outcomes, but also offers an opportunity for professional learning about both the subject and assessment itself. This can ensure teacher assessment is 'reliable enough'. Greater shared understanding will also improve validity. The implication is that assessment must be part of professional learning discussions, and not be considered separately (Earle, 2017).

Formative and summative

In 1989 the British Educational Research Association (BERA) formed the Assessment Reform Group (ARG) as a voluntary group of researchers to ensure assessment policy and practice took account of relevant research. It was dissolved in 2010 after two key studies: *Inside the Black Box* (Black and Wiliam, 1998) and *Beyond the Black Box* (Broadfoot *et al.*, 1999). An additional publication was published independent of the ARG with advice for improving classroom assessment: *Working inside the Black Box* (Black *et al.*, 2004). The project aimed to understand how formative assessment could improve learning in the classroom.

In seeking to define assessment for learning in practice, the ARG identified some key characteristics which it contrasted with those that added tests or procedures, or those that simply reported grades or marks to pupils:

- It is embedded in a view of teaching and learning of which it is an essential part.
- It involves sharing learning goals with pupils.
- It aims to help pupils to know and to recognise the standards they are aiming for.
- It involves pupils in self-assessment.
- It provides feedback which leads to pupils recognising their next steps and how to take them.
- It is underpinned by confidence that every student can improve.
- It involves both teacher and pupils reviewing and reflecting on assessment data (Broadfoot *et al.*, 1999, p. 7).

Earle reminds us that it is important to consider the *use*, rather than the activity in determining whether assessment is formative or summative. Most assessment can be utilised for both purposes (Earle, 2021).

Newton argues that the distinction between the two is 'spurious' and may actually hinder the development of assessment practice. The lack of precision in definitions can lead to confusion, for example in viewing formative assessment as an event rather than assessment with a particular purpose, i.e. as a judgement rather than the use to which an assessment judgement can be put (Newton, 2007b). Work by Brookhart also suggests that students don't make a clear distinction between summative and formative assessment, and higher achieving students also make use of formative assessment even in summative situations (Brookhart, 2001).

Wiliam identifies five key strategies of formative assessment:

1. clarifying, sharing and understanding learning intentions and success criteria
2. eliciting evidence of learning
3. providing feedback that moves learning forward
4. activating learners as instructional resources for one another (cf. peer assessment)
5. activating learners as owners of their own learning (cf. self-assessment)

(Wiliam, 2018, p. 2)

A randomised controlled trial project into the effectiveness of embedding formative assessment by the EEF found students in the schools participating made the equivalent of two additional months' progress in their attainment 8 GCSE score (using EEF's conversion tool). There was greater additional progress for those children in the lowest third for prior attainment. Whilst the formative assessment content was considered similar to existing approaches being implemented in schools, the sustained focus on reinforcing those practices was considered to set the intervention apart (Speckesser *et al.*, 2018).

Aside from the issues raised by Newton and others relating to attempts to distinguish between formative and summative assessment, there has been other criticism of formative assessment, or assessment for learning (AfL). One systematic review argues that the vast majority of AfL studies are small-scale action research designs. Definitions are wide and research designs may lack an action theory or lack systematic data collection. This leads them to conclude that claims relating to the effects of AfL have been over-sold by some authors, though they also acknowledge a modest impact on teaching and learning (Baird *et al.*, 2014).

There is some criticism of formative assessment. In particular, the concerns are that large effect sizes are not replicable, there is under-representation of measurement principles and unclear impact on the education system. Robust empirical studies on formative assessment are lacking in the literature according to Baird *et al.*, 2014.

In *Making Good Progress*, Christodoulou points out that despite decades during which assessment for learning has been a focus of national policy and widely supported by teachers and educators, it has not had the kind of success expected (Christodoulou, 2016). Coe also expresses this belief that AfL has not delivered the impact it promised (Coe, 2013). Much of the criticism is about how the idea has been implemented, but it has also been suggested that government support has been counter-productive and encouraged a summative approach by linking it to monitoring pupils' progress. The two purposes of formative and summative assessment are essentially competing within assessment systems (Christodoulou, 2016).

Early years

The early years foundation stage profile is not mandatory. In the early years assessment is done formatively in the three prime areas of learning: personal, social and emotional development (PSED), communication and language, and physical development. Additional areas of learning are literacy development, mathematics, understanding the world and expressive arts and design. Profile judgements should be made by cumulative observational evidence recorded throughout the year (Standards and Testing Agency, 2020a). The early learning goals that set out the expected level of development attained by the end of the EYFS should be assessed with a

holistic, best-fit judgement based on their own knowledge of the child and own expert professional judgement (DfE, 2021b).

Assessment as proxy for learning

Both descriptor-based and task-based assessment systems describe performance but don't analyse it. Christodoulou identifies a number of flaws with these approaches to measuring progress:

- they expect two different inferences from the same assessment
- can lead to overgrading and overtesting
- unhelpful feedback
- leads to the measurement of formative progress with summative grades
- inadvertently encourages a focus on short-term performance over long-term learning (Christodoulou, 2016).

Whilst assessment can help us focus on learning rather than engagement, it may still be a poor proxy for progress.

Robert Bjork talked about the importance of dissociating performance from learning, with the former something that is easier to measure but can only be used to infer learning indirectly.[1]

High stakes assessment

High stakes assessment has two main potential pitfalls: increased levels of stress and anxiety for both pupils and staff, and pressure to adopt morally dubious practices in a culture of performativity (Meadows and Black, 2018).

In the aftermath of cancelled exams due to COVID-19, there has been increased scrutiny of centre-assessed grades (2020) and teacher-assessed grades (2021), with many accusations of grade inflation. In 2022, Ofqual confirmed some adaptations to make exams fairer for those students who have experienced significant absence from school. Grades will be based on an average taken of 2019 and 2021 with 2023 similar to 2019 – intending to return to the pre-pandemic grade distribution over two cohorts.[2]

Teaching to the test

One of the unintended consequences of assessment can be that of 'teaching to the test'. The teacher narrows their approach to the topics expected to be assessed, or schools restrict their curriculum to core subjects that contribute to accountability measures. Ofsted's research into the curriculum that influenced the new Education Inspection Framework of 2019 (Ofsted, 2019) found that the consequence of over focus on exams led to a narrowing of the curriculum in primary and a reduction of Key Stage 3 in many schools (Spielman, 2018).

As Christodoulou says, exams are only samples from the wider domain, so 'the moment we start to target the exam, then the exam will stop being a valid measure' (Christodoulou, 2016, p. 145). Threats to the validity of the inferences we can make from exams come from actions such as cramming, boosting short-term performance rather than long-term learning. But also focusing heavily on preparing pupils to answer particular types of exam question, which may seem like good practice, actually may compromise the validity of the results. Excessive focus can inflate scores on high-stakes tests that are not reflected in other tests which draw from the same domain. Improvements in exam performance can therefore reflect more successful coaching rather than actual improvement in learning, as this success is not matched elsewhere (Koretz, 2008)

Counsell also talks about the damage of teaching to the test which misses the point of the curriculum entirely and may limit success for many (Counsell, 2018).

[1] Robert Bjork discussing performance and learning posted by gocognitive in 2012 https://www.youtube.com/watch?v=MMixjUDJVIw [accessed 04/03/2022]

[2] https://ffteducationdatalab.org.uk/2021/09/what-impact-will-ofquals-chosen-grading-system-in-2022-have/ [01/03/2022]

Measuring progress

Measuring pupil progress is problematic (CEM, 2019), making it an uncertain method of evidence to measure teacher effectiveness. David Didau challenges the very concept of progress in terms of an implicit belief that results should always be improving, arguing that it is not possible for student learning to progress in a short time, or at a great rate, and continued for an extended period (Didau, 2015).

Learning progressions such as those espoused in KS3 levels or Assessing Pupil Progress (APP) were abandoned as it was recognised that huge lists of descriptors assigned a linear progression which did not match the progression of most students (Ashman, 2019). Evaluating teachers by tracking pupil progress may have a negative impact by distorting pedagogy (Gibbons, 2019) and Ofsted has stated that internal school data will no longer be used as evidence relating to progress (Harford, 2018).

In the absence of levels, many secondary schools have adopted flight paths that reduce GCSE grades to a linear progression. As GCSE grades are summative and norm-referenced, and grounded in the curriculum, applying a grade to a Year 7 or suggesting they move from a grade 2 to 3 in Year 8, just replaces one inaccurate and vague system with another (Ford, 2016a). We need to understand what it means for a student to get better at a subject without the use of meaningless grades (Ford, 2016b), before we can even begin to think about using pupil progress across a one-year period as being a valid means of evaluating teacher effectiveness.

> *A good assessment system must not only clarify the current state and the goal state... but it must also establish a path between the two.*

(Christodoulou, 2016, p. 142)

Defining and communicating the model of progression in different subjects is the first step. Prose statements or exam specifications and past papers are the dominant methods, but they are not specific or detailed enough. An alternative is the textbook which offers curriculum coherence, but this has tended to fall out of favour in England according to Oates in (Christodoulou, 2016).

Developing a progression model requires clarity on the end goal. Exams are only a sample of a wider domain and so the end goal must be wider than this for exams to be a valid measure of the goal of mastery of a particular domain. Research to support the development of progression models exists in some subjects and not others, for example in reading. This enables us to identify the fundamental building blocks which enable further learning. Christodoulou also suggests however that a progression model might be better thought of in terms of important concepts and achievements rather than individual subjects (Christodoulou, 2016).

In terms of measuring progress in learning, this has traditionally been done primarily through ever finer grades, much like a measuring tape to measure height. However, learning isn't linear. A suggested alternative metaphor is a marathon runner where the desired outcome is measured in time, but to improve at running the coach might suggest short runs, or interval training, or even strength training rather than running. This approach in education might be seen through a variety of classroom exercises used to develop skill – used to support rather than measure progress (Christodoulou, 2016).

Curriculum as progress model

Christodoulou highlights the problem of drawing different inferences from the same assessment, or relying on one type of assessment to gives us all the information we need. There is a link between formative and summative assessment in that the formative assessments should assess tasks that support the final performance. This link is described as the model of progression by Wiliam (Christodoulou, 2016).

Christodoulou offers the textbook as a specific and detailed communication of the model of progression. The format creates curriculum coherence, but they have fallen out of use in England to a lesser or greater degree. She suggests we begin with the curriculum aims when planning a progression model so that we focus on teaching to the domain rather than the exam (Christodoulou, 2016).

The curriculum as progress model tells us how well students have learned a particular aspect of the curriculum. According to Didau, it should be seen primarily as a statement of competence rather than be used to discriminate between students to rank or assign summative statements of progression. This means that over the long term, student performance is a good tool to enable us to consider the design of the curriculum or its teaching. For example, if the majority fail to reach the threshold test score (say 80%), then this would suggest a problem with the instruction or content design. What we can't or shouldn't do with this model is compare the % achieved by a student in term 1, with the % achieved in term 6 – we cannot measure progress, only performance. We can compare lateral performance but not longitudinal.

> *Tests are very useful for assessing how well students have learned particular curriculum content, but cannot be used to measure the rate at which students are progressing towards better future test performance.*

(Didau, 2021)

Methods of assessment

Exam-based assessment

Exams assess a sample of the domain. Currently this is the national standardised system used at the end of KS2 (Year 6 SATs), at the end of KS4 (GCSEs/BTECs), and KS5 (A-levels, BTECs). This model works by isolating the task that is to be used to make the summative inference. Exams based on a difficulty model, where questions increase in difficulty, can generate a significant level of detail at question level – this is often referred to as question-level analysis (QLA). Exams based on the quality model are more difficult as they generally have fewer questions and require marker judgement supported by rubric. This means that exams in different subjects may vary significantly (Christodoulou, 2016).

Rubrics

Rubrics are descriptive statements that articulate the expectations for an assignment or task. They list assessment criteria and describe levels of quality or marks in relation to each of these criteria.

The use of rubrics in summative assessment has received more attention than for formative purposes. Research suggests that rubrics have the potential to enhance student performance but it remains inconclusive, with mixed results. When used by teachers, rubrics can enhance the alignment of assessment and instruction. In a student-centred approach, at least one study demonstrated significant positive impacts when students used rubrics for self-assessment (Panadero and Jonsson, 2013).

Some of the ways in which the use of rubrics may facilitate student performance:

- increased transparency in terms of expectations
- reduced anxiety around assessment
- aid the feedback process
- improved student self-efficacy (e.g. generating criteria for a model essay and using it for self-assessment of drafts)

However, other factors moderate these effects. Rubrics may be combined with other instructional interventions, such as metacognitive activities. Results are more positive in older students and therefore longer and larger interventions are needed in schools to produce similar results (Panadero and Jonsson, 2013).

Multiple choice questions (MCQ)

MCQs offer a low stake, effective means of retrieval practice. An experimental study by Smith and Karpicke (2014) indicated that there is little advantage in answering short-answer questions over MCQs in terms of achieving the best learning (Smith and Karpicke, 2014).

This study considers the risks of 'lures', where students may learn false facts from MCQ tests. However, the positive effects outweigh this cost, with the benefits to performance being not simply in simple definitional questions, but also for higher level concept questions (Marsh *et al.*, 2007).

Paul Moss draws on the research to outline the key characteristics of effective use of MCQs:

1. Build the level of difficulty gradually – understand the schema.
2. Master specific knowledge first.
3. Actively engage retrieval, e.g. avoiding 'none of the above', including at least 2 plausible answers.
4. Mitigate guessing, e.g. asking several questions about the same topic or having 4 options.
5. Include mastery pathways (most useful online where error can lead to a different set of questions to address the gap in core knowledge) (Moss, 2020).

A further consideration is the use of weighted MCQs. These ask test-takers to indicate their level of confidence in the correctness of one alternative compared with the others. In experiments, this confidence-weighted approach led to greater benefits in the ability of test-takers to answer new but related questions (Sparck, Ligon Bjork and Bjork, 2016).

Confidence in understanding is an important element in learning and self-assessment of confidence. MCQ answers can lead to improved performance as the format encourages considered rejection and selection of answers rather than guessing.

In summary:

- Confidence assessments can help both teachers and pupils in the learning process.
- Giving a confidence rating on individual questions can aid retention and improve performance, as well as incentivising knowing-not-guessing.
- The current evidence that confidence assessments are effective is mostly limited to multiple-choice tests.
- Confidence assessments are most effective when completed privately (Cambridge Mathematics, 2016).

Cynthia Brame at Vanderbilt University authored a guide to writing good multiple choice questions based on the research by Haladyna, Downing and Rodriguez, 2010:

- constructing an effective stem
- constructing effective alternatives
- additional guidelines

(Brame, 2013)

Quizzing

The '**testing effect'** is where long-term retention in the memory results from taking a memory test. Roediger and Karpicke's experiments demonstrated that this testing effect was not a result of an opportunity to re-study material and that prior testing produced greater retention than studying on delayed tests. Whilst students felt more confidence in repeated studying, testing is a more powerful means of improving learning and not just assessing it (Roediger and Karpicke, 2006).

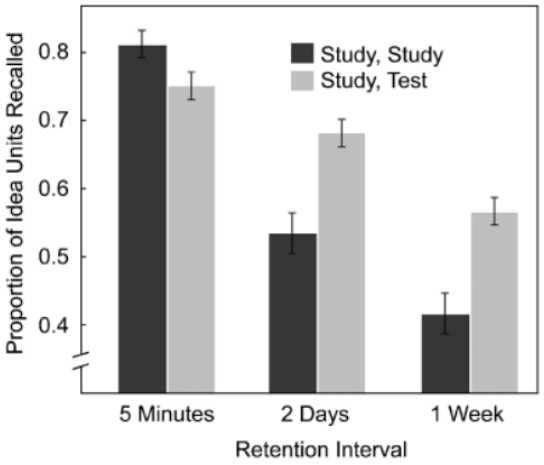

Fig. 1. Mean proportion of idea units recalled on the final test after a 5-min, 2-day, or 1-week retention interval as a function of learning condition (additional studying vs. initial testing) in Experiment 1. Error bars represent standard errors of the means.

(source: Roediger and Karpicke, 2006, p.250)

There is lots of evidence that testing previously studied information enhances long-term memory. However, there is also some research that suggests **pretesting** enhances learning, even where initial retrieval is unsuccessful. In this manner, testing or quizzing can be seen as a learning event rather than solely a means of assessment. This direct impact on memory from pretesting is in addition to affecting the attention and intentions of learners (Richland *et al.*, 2009).

Yang et al. looked at the factors that modulated the magnitude of the effects found from quizzing on student academic achievement. Their meta-analysis found that testing had an advantage over other learning strategies for learning factual knowledge, concept comprehension and knowledge application. Overall they suggest testing is not only an assessment of learning but also for learning (Yang *et al.*, 2020).

High confidence errors are actually more easily corrected than low confidence errors, a phenomenon known as **hypercorrection**. However, a delay in testing may result in these errors recurring. Testing immediately after corrective feedback can enhance memory for not only the correct answers, but also block the return of errors (Metcalfe and Miele, 2014).

Further work on retrieval practice suggests that it offers benefits for memory precision, not just

discrete right-or-wrong answers (Schuetze, Eglington and Kang, 2019).

A well-designed test will have items that get progressively more difficult, giving a better measure of individual students' performances. So if one student achieves 70% on a test, and another achieves 35%, this does not mean the first student has done twice as well. It is far easier to get less than 35% than it is to get more than 70%. To establish item difficulty, you can see which questions are answered correctly by more students – thereby indicating that it is an easier question (Didau, 2021).

Peer and self-assessment

One meta-analysis of peer assessment found mixed results on student learning, despite a prevailing positive view. Their findings suggested a positive increase in performance for those students participating in peer assessment but that this was substantially larger when the students received rater training (Li *et al.*, 2020).

Another meta-analysis of experimental and quasi-experimental studies found an overall small to medium effect of peer assessment on academic performance. This suggests it improves performance compared with no assessment or teacher assessment, but not significantly differently from self-assessment. Importantly, they found that performing both the role of assessor and being assessed themselves, potentially benefitted students' learning more than just being assessed. However, the efficacy of peer assessment is likely to be modified by factors relating to the student and the assessment itself, as well as the nature of the learning environment (Double, McGrane and Hopfenbeck, 2020).

An integrated assessment system

Christodoulou considers how to improve assessments by bringing together different considerations around summative and formative assessment and progression models. She argues that an accurate and useful progression model is the foundation of any assessment system because it explains how students improve. She suggests the most effective method of bringing the required resources together is a textbook. The system would include formative and summative item banks to provide coherence, pupil ownership, and data that supports a self-improving system (Christodoulou, 2016).

Formative feedback

Recognition of the importance of feedback for a time resulted in fashionable practices such as deep written marking and triple marking. Unsurprisingly this led to a frightening workload burden without any clear evidence of impact on learning, and the beginning of a new debate around effective approaches to assessment, marking and feedback (DfE, 2016a).

Feedback done well can support pupil progress and address misconceptions. However, done poorly, it can even harm progress. Getting feedback right is therefore crucial. Part of the focus historically has been on how feedback is delivered – debates between written or verbal – rather than the principles of effective feedback.

> *Feedback has to be part of a system that is set up in such a way that the information can actually be used to improve it.*

(Dylan Wiliam, EEF, 2021, p. 5)

As with many aspects of teaching, a teacher's engagement feedback comes with an 'opportunity cost' – it can take up a large amount of teacher time, thereby reducing the time they can give to other tasks (EEF, 2021).

Hattie and Timperley's work on feedback emphasises its powerful influence, both positive and negative. The type of feedback and how it is given makes a significant difference to its effectiveness. They focus on three important questions: where am I going? How am I going? and Where to next? A key theme is ensuring that feedback is targeted at the appropriate level for the student in order to reduce the discrepancy between current understanding and the desired level of understanding (Hattie and Timperley, 2007).

Another important element of feedback is ensuring that the student trusts the feedback given. They describe 'wise feedback' as critical feedback that assuages mistrust by emphasising two things: the high standards expected by the teacher alongside the belief that the student was capable of meeting those standards (Yeager *et al.*, 2014).

Written vs verbal

An EEF review of the evidence on written marking highlights how this central feature of teaching has long driven high teacher workloads, despite a lack of high-quality evidence of its efficacy. The findings they highlight suggest that:

- Careless mistakes are best marked as incorrect, without giving the correction.
- Errors resulting from misunderstanding are best addressed with hints and questions that lead to underlying principles.
- The use of specific and actionable targets is most likely to increase pupil progress.
- Some forms of marking, e.g. acknowledgement marking, are unlikely to enhance progress (Elliott *et al*., 2016).

A small scale Randomised Control Trial (RCT) carried out in three secondary schools to replace written feedback with alternative approaches found a positive effect on the reduction of teacher workload and perception of their work, a sense of frustration amongst students, and no detectable impact (positive or negative) on student outcomes (Kime, 2018).

An action research project carried out by teachers in conjunction with UCL also explored verbal feedback as an alternative to written marking (McGill and Quinn, 2019).

Grades

Butler (1988) claims that awarding grades skews interest away from learning and progress. She found that grades and grades with comments generally undermined both interest and performance although high achievers were likely to maintain their interest and motivation in anticipation of further grades (Butler, 1988).

A systematic review looking at the impact of summative assessment and tests on students' motivation for learning suggests that high stakes tests lower the self-esteem of low-achieving pupils, an effect reinforced by repeated testing (Assessment and Learning Research Synthesis Group, 2002).

A study from Sweden found that the introduction of grades and increased testing increased school-related stress and reduced academic self-esteem. It also had an indirect effect on psychosomatic symptoms and life satisfaction. The negative effects were stronger for girls (Högberg *et al.*, 2021).

Comparative judgement

Research suggests that it is possible to judge pupil responses accurately and quickly through a process known as comparative judgement. This approach is growing in interest due to the reduction of workload for teachers without loss of validity if well implemented.

Some of the key advantages of comparative judgement:

- greater reliability
- efficiency
- reduces bias (through anonymising responses and multiple viewings by multiple assessors)
- robust data trail

(No More Marking, 2020)

Jones and Wheadon explored the use of comparative judgement for peer assessment. Comparative judgement does not require explicit, detailed criteria and therefore offers advantages for peer assessment in a number of ways: firstly for judging 'creativity' and other skills not easily operationalised in rubrics; secondly for unpredictable responses which are hard to anticipate in rubrics; and finally it doesn't require training in the way that rubric judgements do (Jones and Wheadon, 2015).

Technology

A number of methods of digital or online assessment exist currently, from self-marking quizzes on platforms such as Google classroom or Microsoft forms, to more comprehensive assessment packages such as Cognitive Ability Tests (CATs). Some of these technologies reduce workload for teachers and have specific benefits for students, such as supporting retrieval practice. However, others require more research or are more controversial.

Algorithms

The use of algorithms has been explored to improve the consistency of marking essays in humanities subjects. This involved identifying indicators that account for scores and turning them into an algorithm that can be used to assess new essays. Unfortunately, this creates a risk that if the teachers and students know that the algorithm rewards length for example, they are encouraged to write more at the potential expense of quality. Studies demonstrated a number of ways that this system could be 'gamed', for example by repeating paragraphs (Christodoulou, 2020). More recently the proposed use of an algorithm to award grades in Summer 2020 exams generated significant issues. Concerns regarding algorithms may raise an interesting paradox, according to one study, as people believe that they understand human decision making better than algorithmic decision making, despite this understanding being illusory. In fact, we don't understand human decision-making processes (Bonezzi, Ostinelli and Melzner, 2022).

Accountability

Most of the variation between student scores occurs within schools, rather than between schools. The use of standardised tests in high-stakes accountability regimes may therefore have shortcomings that undermine the interpretations routinely drawn from the grades generated. The key assumption in using testing for accountability is that student outcomes can be attributed primarily to differences in the quality of education received. However, only a small proportion of variation in outcomes (as little as 10%) can be attributed to the quality of schooling, making inferences from such scores about the quality of education problematic. Wiliam recounts a long history of concerns relating to high stakes accountability measures and how these may incentivise narrower or mechanistic teaching, or even dishonesty. Further research in the US revealed that between-school variance was largely accounted for by differences in the social class of individual students, rather than differences in school quality. In England, value-added measures were introduced to try to mitigate the influence of this factor, at a school-level rather than individual teacher-level. (Wiliam, 2010).

Wiliam explores the nature of testing in detail to support the conclusion that the progress of individual students is slow compared to the variability of achievement within the age cohort. As a result it is factors over which schools have little influence, for example prior achievement, that determine variance in students' scores rather than the quality of the education provided by the school. Therefore such test scores do not successfully support inferences about the quality of education provided by a school (Wiliam, 2010).

Nevertheless, there is evidence that when test results as a performance indicator are the focus of public policy, then performance (as measured by that test) improves – an effect known as Goodhart's Law (Kellner, 1997 in (Wiliam, 2010)).

NFER's review of accountability gives international comparisons. Some of the findings echo those from elsewhere including the use of pupil performance as a high stakes accountability measure leading to privileging certain aspects of the curriculum over others, or 'teaching to the test'. The application of accountability measures can lead to an increase in the achievement gap through, for example, an over-focus on the performance of 'borderline' pupils, though they can also be used to decrease the gap (Brill *et al.*, 2018).

Secondary schools

Accountability measures 2019

- progress 8
- EBacc entry
- pupil destinations
- attainment in English and mathematics
- attainment 8
- EBacc APS (average point score)

The DfE states that the aim of these measures is to encourage schools to offer a broad and balanced curriculum with a focus on an academic core at KS4, and to reward schools for the teaching of all their pupils (DfE, 2020b)

Attainment and progress 8

Progress 8 was introduced as a performance measure in 2016 to capture the progress a student makes from the end of primary school (KS2) to the end of secondary school (KS4/GCSE). It is based upon an average of maths and English at KS2 and individual points are used to calculate a school's score (DfE, 2016c). As a measure of total achievement across all subjects, KS2 can be a good indicator of GCSE success, but it is less convincing at subject level (Benton and Sutch, 2014). More importantly for schools, progress scores are not directly comparable from year to year, only a change in progress banding indicates a change in performance (DfE, 2020a). If you combine this with the evidence pointing to the variable probability of receiving the 'definitive' mark and therefore grade at GCSE across difference subjects (0.96 in mathematics but only 0.52 in English language and literature) it is clear that progress 8 is a problematic measure at best (Black, Rhead and Pinot de Moira, 2018). Progress 8 is also highly influenced by the percentage of EAL students whose progress is often underestimated as a result of KS2

assessments being taken before they reach fluency in English (Hazel, 2018).

Progress 8 is a measure of attainment net of the effect of prior attainment – it is not a measure of school effectiveness. Value added models of attainment, such as Progress 8, tend to show that differences in attainment between schools are small. There is substantial variation in pupil performance in all schools, with relatively little being between schools (FFT Education Datalab, 2021).

National reference tests

National reference tests are designed to monitor pupil performance over time and inform GCSE grades in English and maths. The 2021 tests show that maths performance is closer to 2017 when GCSEs were reformed, but English performance shows no statistical difference (Ofqual, 2021).

Primary schools

Year 6 progress measures

Progress measures introduced in 2016 seek to compare pupil results to those of pupils nationally with similar prior attainment. Much like P8, it is intended to reflect a value-added dimension to the measure of how well a school is doing. They are therefore intended to be fairer to schools in challenging circumstances as they recognise both the start and end point. The progress measure is based on an average score based on performance in KS1 across reading, writing and mathematics weighted 50:50 for English and maths. The school level progress score is calculated from an average of the individual progress scores of each pupil in Year 6. Confidence intervals are given to account for the uncertainty of the effectiveness of a school based on a single cohort of pupils (DfE, 2016b). Technical guidance is available (DfE, 2019).

In KS2 for the academic year 2021/2022, tests were timetabled from 9 May to 12 May 2022.[3]

KS1 assessments

National curriculum assessments at KS1 were cancelled for the 2020/2021 academic year. Guidance for the testing periods in 2021/2022 was published in October 2021 (Standards and Testing Agency, 2021).

Teachers must assess English reading, writing and mathematics, for those pupils that have completed the KS1 programmes of study and are working at the standard of national curriculum assessment, using the TA frameworks. For those working below this standard, the pre-key stage standards should be used.

Teacher assessments must be based on sound and demonstrable evidence as well as the teacher's knowledge of pupils. (Standards and Testing Agency, 2020c).

Phonics assessments

The phonics check is designed to confirm that a pupil has learnt phonic decoding. It consists of twenty real words, and twenty pseudo words that pupils read aloud. The test is taken by pupils who will be age 6 by the end of the academic year (Year 1). It is also taken by those who will be age 7 (Year 2) if they did not meet the expected standard previously, with some specific exemptions. School-level results are not published (Standards and Testing Agency, 2021).

Early years

In the early years foundation stage (EYFS), there are three statutory assessments, the progress check at age 2, the reception baseline assessment and the early years foundation stage profile. This latter assessment is a summative assessment designed to check they are progressing well and meeting national requirements and provides a report sent home to parents.

[3] https://www.gov.uk/guidance/primary-assessments-future-dates

Reception baseline

The new reception baseline assessment became statutory in September 2021. Schools must administer the reception baseline assessment for each child in the first 6 weeks after they enter reception.

The purpose of the assessment is to provide on-entry assessment of pupil attainment as a starting point in order to generate a cohort-level progress measure to the end of KS2. It is not intended to provide formative information for practitioners, to be used to measure performance in early years, or provide diagnostic information about pupils' areas for development.

The assessment consists of age appropriate tasks in mathematics, literacy, communication and language (Standards and Testing Agency, 2020b).

Further statutory guidance is available (DfE, 2021a).

COVID-19 note

Due to the cancellation of statutory KS1, KS2, GCSE, AS, A-level, and other vocational and technical qualifications in 2020 and 2021, the DfE has announced that grades based on alternative assessment arrangements will not be used to produce the usual performance measures.[4]

[4] https://www.gov.uk/government/publications/coronavirus-covid-19-school-and-college-performance-measures/coronavirus-covid-19-school-and-college-accountability-2020-to-2021 [accessed 26/01/2022]

References

Ashman, G. (2019) *Learning progressions are invalid and inequitable | Filling the pail*, *gregashman*. Available at: https://gregashman.wordpress.com/2019/12/13/learning-progressions-are-invalid-and-inequitable/ (Accessed: 1 December 2020).

Assessment and Learning Research Synthesis Group (2002) *A systematic review of the impact of summative assessment and tests on students' motivation for learning*, *EPPI Review*. Available at: https://dspace.stir.ac.uk/bitstream/1893/19607/1/SysRevImpSummativeAssessment2002.pdf (Accessed: 4 March 2022).

Baird, J.-A. *et al.* (2014) *Assessment and Learning: State of the Field Review.* Available at: https://www.researchgate.net/publication/263654863_Assessment_and_Learning_State_of_the_Field_Review (Accessed: 6 January 2022).

Benton, T. and Sutch, T. (2014) *Analysis of use of KS2 data in GCSE predictions*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/429074/2014-06-16-analysis-of-use-of-key-stage-2-data-in-gcse-predictions.pdf (Accessed: 3 March 2020).

Black, B., Rhead, S. and Pinot de Moira, A. (2018) *Marking consistency metrics An update*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf (Accessed: 3 March 2020).

Black, P. *et al.* (2004) 'Working inside the black box: Assessment for learning in the classroom', *Phi Delta Kappan*. Phi Delta Kappa Inc., 86(1), pp. 8–21. doi: 10.1177/003172170408600105.

Black, P. and Wiliam, D. (1998) 'Inside the Black Box: Raising Standards Through Classroom Assessment', *Phi Delta Kappan*, 80(2), pp. 139–148. doi: 10.1002/hrm.

Bonezzi, A., Ostinelli, M. and Melzner, J. (2022) 'The human black-box: The illusion of understanding human better than algorithmic decision-making.', *Journal of Experimental Psychology: General*. American Psychological Association (APA). doi: 10.1037/XGE0001181.

Brame, C. (2013) *Writing Good Multiple Choice Test Questions*, *Center for Teaching*. Available at: https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/ (Accessed: 20 May 2021).

Brill, F. *et al.* (2018) *What Impact Does Accountability Have On Curriculum, Standards and Engagement In Education? A Literature Review*. Available at: https://www.nfer.ac.uk/media/3032/nfer_accountability_literature_review_2018.pdf (Accessed: 23 March 2022).

Broadfoot, P. *et al.* (1999) *Beyond the black box*. Available at: https://www.nuffieldfoundation.org/sites/default/files/files/beyond_blackbox.pdf.

Brookhart, S. M. (2001) 'Successful Students' Formative and Summative Uses of Assessment Information', *Assessment in Education: Principles, Policy & Practice*. Taylor & Francis Group, 8(2), pp. 153–169. doi: 10.1080/09695940123775.

Business Insider (2014) 'Interview: Amazon CEO Jeff Bezos'. Available at: https://www.youtube.com/watch?v=Xx92bUw7WX8 (Accessed: 11 September 2020).

Butler, R. (1988) 'ENHANCING AND UNDERMINING INTRINSIC MOTIVATION: THE EFFECTS OF TASK-INVOLVING AND EGO-INVOLVING EVALUATION ON INTEREST AND PERFORMANCE', *British Journal of Educational Psychology*. John Wiley & Sons, Ltd, 58(1), pp. 1–14. doi: 10.1111/J.2044-8279.1988.TB00874.X.

Cambridge Mathematics (2016) ‘How does assessing confidence affect learning and testing in mathematics?’, *Espresso: research filtered by Cambridge Mathematics*, (2). Available at: https://www.cambridgemaths.org/Images/espresso_2_confidence_assessments_in_mathematics_learning.pdf (Accessed: 20 May 2021).

CEM (2019) *Measuring Progress in Education*, *CEMblog*. Available at: http://www.cem.org/blog/measuring-progress-in-education/ (Accessed: 1 December 2020).

Christodoulou, D. (2016) *Making good Progress? The future of assessment for learning*. Oxford University Press (OUP).

Christodoulou, D. (2020) *Teachers vs Tech? The case for an ed tech revolution*. Oxford: Oxford University Press (OUP).

Coe, R. (2013) *Improving Education: A triumph of hope over experience Inaugural Lecture of Professor*. Available at: http://www.cem.org/attachments/publications/ImprovingEducation2013.pdf (Accessed: 6 December 2019).

Counsell, C. (2018) *Senior Curriculum Leadership 1: The indirect manifestation of knowledge: (B) final performance as deceiver and guide | the dignity of the thing*. Available at: https://thedignityofthethingblog.wordpress.com/2018/04/12/senior-curriculum-leadership-1-the-indirect-manifestation-of-knowledge-b-final-performance-as-deceiver-and-guide/ (Accessed: 14 September 2020).

DfE (2016a) *Eliminating unnecessary workload around marking*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/511256/Eliminating-unnecessary-workload-around-marking.pdf (Accessed: 19 April 2021).

DfE (2016b) *Primary progress measures How the primary progress measures are calculated*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/560969/Primary_school_accountability_summary.pdf.pdf (Accessed: 20 May 2021).

DfE (2016c) *Progress 8 How Progress 8 and Attainment 8 measures are calculated*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/561021/Progress_8_and_Attainment_8_how_measures_are_calculated.pdf (Accessed: 3 March 2020).

DfE (2019) *Primary school accountability in 2019: technical guide A technical guide for primary maintained schools, academies and free schools*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/854515/Primary_school_accountability_in_2019_technical_guide_2_Dec_2019.pdf (Accessed: 20 May 2021).

DfE (2020a) *Progress scores for key stage 4: school and college performance tables*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/863557/Key_stage_4_progress_banding_calculations_bandings_2019.pdf (Accessed: 3 March 2020).

DfE (2020b) *Secondary accountability measures Guide for maintained secondary schools, academies and free schools*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/872997/Secondary_accountability_measures_guidance_February_2020_3.pdf (Accessed: 20 May 2021).

DfE (2021a) *Reception baseline assessment and reporting arrangements*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/983889/Reception_baseline_assessment_and_reporting_arrangements_v1.0.pdf (Accessed: 1 July 2021).

DfE (2021b) *Statutory framework for the early years foundation stage Setting the standards for learning, development and care for children from birth to five*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974907/EYFS_framework_-_March_2021.pdf (Accessed: 24 May 2021).

Didau, D. (2015) *The myth of progress*, *learningspy.co.uk*. Available at: https://learningspy.co.uk/featured/the-myth-of-progress/ (Accessed: 1 December 2020).

Didau, D. (2021) *Why using the curriculum as your progression model is incompatible with 'measuring progress'*, *learningspy.co.uk*. Available at: https://learningspy.co.uk/assessment/why-using-the-curriculum-as-your-progression-model-means-you-cant-measure-progress/ (Accessed: 3 December 2021).

Double, K. S., McGrane, J. A. and Hopfenbeck, T. N. (2020) 'The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies', *Educational Psychology Review*. Springer, 32(2), pp. 481–509. doi: 10.1007/S10648-019-09510-3/FIGURES/4.

Earle, S. (2017) '' "But I've not got time for any more assessment": balancing the demands of validity and reliability'', *Impact*, 1, pp. 44–46. Available at: http://researchspace.bathspa.ac.uk/10152/1/10152.pdf.

Earle, S. (2021) 'Principles and purposes of assessment in the classroom', *Impact*, (12), pp. 20–23. Available at: http://researchspace.bathspa.ac.uk/14169/1/14169.pdf (Accessed: 6 January 2022).

EEF (2021) *Teacher Feedback to Improve Pupil Learning | Education Endowment Foundation | EEF*. Available at: https://educationendowmentfoundation.org.uk/public/files/Publications/Feedback/Teacher_Feedback_to_Improve_Pupil_Learning.pdf (Accessed: 14 June 2021).

FFT Education Datalab (2021) *All models are wrong 1 , some of these might be useful: Options for adjusting school performance indicators for context Acknowledgments Background*. Available at: https://mk0fftеducation79fru.kinstacdn.com/wp-content/uploads/2021/03/cva_report_published.pdf (Accessed: 26 March 2021).

Ford, A. (2016a) *Creating flight paths to replace levels Year 7-11 - the impact of the new GCSE grade descriptors*, *AndAllThat blog*. Available at: http://www.andallthat.co.uk/blog/creating-flight-paths-to-replace-levels-year-7-11-the-impact-of-the-new-gcse-grade-descriptors (Accessed: 1 December 2020).

Ford, A. (2016b) *Shackled to a Corpse? Why can't we make progress in our understanding of progression?*, *AndAllThat blog*. Available at: http://www.andallthat.co.uk/blog/shackled-to-a-corpse-why-cant-we-make-progress-in-our-understanding-of-progression (Accessed: 1 December 2020).

Gibbons, A. (2019) *'Early years learning harmed by progress obsession'* | *Tes*, *TES*. Available at: https://www.tes.com/news/early-years-learning-harmed-progress-obsession (Accessed: 1 December 2020).

Haladyna, T. M., Downing, S. M. and Rodriguez, M. C. (2010) 'A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment', *Applied Measurement in Education*. Lawrence Erlbaum Associates, Inc., 15(3), pp. 309–333. doi: 10.1207/S15324818AME1503_5.

Harford, S. (2018) *Assessment – what are inspectors looking at?*, *Ofsted blog*. Available at: https://educationinspection.blog.gov.uk/2018/04/23/assessment-what-are-inspectors-looking-at/ (Accessed: 1 December 2020).

Hattie, J. and Timperley, H. (2007) 'The power of feedback', *Review of Educational Research*, 77(1), pp. 81–112. doi: 10.3102/003465430298487.

Hazel, W. (2018) 'Exclusive: Progress 8 "penalises schools in white working class communities", study shows', *TES news*, April. Available at: https://www.tes.com/news/exclusive-progress-8-penalises-schools-white-working-class-communities-study-shows (Accessed: 19 April 2021).

Högberg, B. *et al.* (2021) 'Consequences of school grading systems on adolescent health: evidence from a Swedish school reform', *Journal of Education Policy*. Routledge, 36(1), pp. 84–106. doi: 10.1080/02680939.2019.1686540/SUPPL_FILE/TEDP_A_1686540_SM5301.DOCX.

James, M. (2006) ‘Assessment, Teaching and Theories of Learning’, in Gardner, J. (ed.) *Assessment and Learning*. Sage, pp. 47–60. Available at: https://www.researchgate.net/publication/271964452_Assessment_Teaching_and_Theories_of_Learning (Accessed: 14 June 2021).

Jones, I. and Wheadon, C. (2015) ‘Peer assessment using comparative and absolute judgement’, *Studies in Educational Evaluation*. Pergamon, 47, pp. 93–101. doi: 10.1016/J.STUEDUC.2015.09.004.

Kime, S. (2017a) *Four Pillars of Assessment: Purpose*, *Evidence Based Education*. Available at: https://evidencebased.education/pillars-assessment-purpose/ (Accessed: 19 April 2021).

Kime, S. (2017b) *Four Pillars of Assessment: Reliability*, *Evidence Based Education*. Available at: https://evidencebased.education/pillars-assessment-reliability/ (Accessed: 19 April 2021).

Kime, S. (2017c) *Four Pillars of Assessment: Validity*, *Evidence Based Education*. Available at: https://evidencebased.education/pillars-assessment-purpose-validity/ (Accessed: 19 April 2021).

Kime, S. (2017d) *Four Pillars of Assessment: Value*, *Evidence Based Education*. Available at: https://evidencebased.education/pillars-assessment-value/ (Accessed: 19 April 2021).

Kime, S. (2018) *Reducing teacher workload: the ‘Re-balancing Feedback’ trial*. Available at: https://dera.ioe.ac.uk/31210/1/Cheshire_Vale_-_Reducing_teacher_workload.pdf (Accessed: 19 April 2021).

Koretz, D. (2008) *Measuring Up: What Educational Testing Really Tells Us*. Harvard: Harvard University Press.

Li, H. *et al.* (2020) ‘Does peer assessment promote student learning? A meta-analysis’, *Assessment and Evaluation in Higher Education*. Routledge, 45(2), pp. 193–211. doi: 10.1080/02602938.2019.1620679.

Marsh, E. J. *et al.* (2007) ‘The memorial consequences of multiple-choice testing’, *Psychonomic Bulletin & Review*, 14(2), pp. 194–199. Available at: https://link.springer.com/content/pdf/10.3758/BF03194051.pdf (Accessed: 20 May 2021).

McGill, R. and Quinn, M. (2019) *UCL Verbal Feedback Project Report 2019*. Available at: https://www.ucl.ac.uk/widening-participation/sites/widening-participation/files/2019_verbal_feedback_project_final_4_print.pdf (Accessed: 10 August 2020).

Meadows, M. and Black, B. (2018) ‘Teachers’ experience of and attitudes toward activities to maximise qualification results in England’, *Oxford Review of Education*. Routledge, 44(5), pp. 563–580. doi: 10.1080/03054985.2018.1500355.

Metcalfe, J. and Miele, D. B. (2014) ‘Hypercorrection of high confidence errors: Prior testing both enhances delayed performance and blocks the return of the errors’, *Journal of Applied Research in Memory and Cognition*, 3, pp. 189–197. doi: 10.1016/j.jarmac.2014.04.001.

Newton, P. E. (2007) ‘Clarifying the purposes of educational assessment’, *Assessment in Education: Principles, Policy and Practice*, 14(2), pp. 149–170. doi: 10.1080/09695940701478321.

No More Marking (2020) *The challenges of assessing writing*. Available at: https://nmm-v2.s3-eu-west-1.amazonaws.com/reports/NMM-NESTA_research_summary.pdf (Accessed: 14 September 2020).

Ofqual (2021) *Ofqual publishes NRT results and contextual analysis 2021*, *gov.uk*. Available at: https://www.gov.uk/government/news/ofqual-publishes-nrt-results-and-contextual-analysis-2021 (Accessed: 10 December 2021).

Ofsted (2019) *Education inspection framework for September 2019*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/801429/E

ducation_inspection_framework.pdf (Accessed: 9 September 2020).

Panadero, E. and Jonsson, A. (2013) 'The use of scoring rubrics for formative assessment purposes revisited: A review', *Educational Research Review*, 9, pp. 129–144. doi: 10.1016/J.EDUREV.2013.01.002.

Richland, L. E. *et al.* (2009) 'The Pretesting Effect: Do Unsuccessful Retrieval Attempts Enhance Learning?', *Journal of experimental psychology*, 15(3), pp. 243–257. doi: 10.1037/a0016496.

Roediger, H. L. and Karpicke, J. D. (2006) 'Test-enhanced learning: Taking memory tests improves long-term retention', *Psychological Science*, 17(3), pp. 249–255. doi: 10.1111/j.1467-9280.2006.01693.x.

Schuetze, B. A., Eglington, L. G. and Kang, S. H. K. (2019) 'Retrieval practice benefits memory precision', *Memory*. Taylor & Francis, 27(8), pp. 1091–1098. doi: 10.1080/09658211.2019.1623260.

Smith, M. A. and Karpicke, J. D. (2014) 'Memory Retrieval practice with short-answer, multiple-choice, and hybrid tests', *Memory*, 22(7), pp. 784–802. doi: 10.1080/09658211.2013.831454.

Sparck, E. M., Ligon Bjork, E. and Bjork, R. A. (2016) 'On the learning benefits of confidence-weighted testing', *Cognitive Research: Principles and Implications*, 1(3). doi: 10.1186/s41235-016-0003-x.

Speckesser, S. *et al.* (2018) *Embedding Formative Assessment Evaluation report and executive summary*. Available at: https://educationendowmentfoundation.org.uk/public/files/EFA_evaluation_report.pdf.

Spielman, A. (2018) *HMCI commentary: curriculum and the new education inspection framework - GOV.UK*. Available at: https://www.gov.uk/government/speeches/hmci-commentary-curriculum-and-the-new-education-inspection-framework (Accessed: 14 February 2020).

Standards and Testing Agency (2020a) *2021 Early years foundation stage assessment and reporting arrangements*. Available at: https://www.gov.uk/government/publications/2021-early-years-foundation-stage-assessment-and-reporting-arrangements-ara (Accessed: 1 March 2022).

Standards and Testing Agency (2020b) *Assessment framework Reception Baseline Assessment*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868099/2020_Assessment_Framework_Reception_Baseline_Assessment.pdf (Accessed: 10 September 2020).

Standards and Testing Agency (2020c) *Key stage 1 teacher assessment guidance For schools and local authorities*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/940852/2021_KS1_teacher_assessment_guidance_V1.0.pdf (Accessed: 1 July 2021).

Standards and Testing Agency (2021) *2022 key stage 1 assessment and reporting arrangements*.

Wiliam, D. (2010) 'Standardized Testing and School Accountability', *Educational Psychologist*, 45(2), pp. 107–122. doi: 10.1080/00461521003703060.

Wiliam, D. (2018) *Embedded formative assessment*. 2nd edn. Bloomington: Solution Tree Press.

Yang, C. *et al.* (2020) 'Testing (Quizzing) Boosts Classroom Learning: A Systematic and Meta-Analytic Review', *Psychological Bulletin*, p. 37. doi: 10.1037/bul0000309.

Yeager, D. *et al.* (2014) 'Breaking the Cycle of Mistrust: Wise Interventions to Provide Critical Feedback Across the Racial Divide', *Journal of Experimental Psychology*, 143(2), pp. 804–824. doi: 10.1037/a0033906.