

What's the wisdom on...



history assessment?

The purpose of this guide

This short guide provides new history teachers with an overview of the 'story so far' of many years of practice-based professional thinking about a particular aspect of history teaching. It draws on tried and tested approaches arising from teachers with many years of experimenting, researching, practising, writing and debating their classroom experience. It therefore synthesises core messages from key *Teaching History* articles, blogs and other publications. The guide includes a range of practical planning suggestions suitable for any key stage and signposts the basic reading essentials for new professionals.

A cautionary tale

Between 1991 and 1995, secondary history teachers in England and Wales had something of a collective awakening about assessment. It followed a huge policy shift in history education: history's first National Curriculum, rolled out in 1991.

History's NC had three 'Attainment Targets', each with 10 levels. Each level was a 'statement of attainment' in disciplinary thinking or skill. Attainment Target 1 went on for ever. It had three strands – causation, change and similarity/difference – resulting in 30 separate statements of attainment. Attainment Target 2 was interpretations of history; Attainment Target 3 was sources. Substantive knowledge was prescribed in the programmes of study but not mentioned in the statements of attainment.

It soon all went horribly wrong. But analysis at the time was swift and sharp, so much so that a big policy shift occurred by 1995. So what did they realise had gone wrong?

History teachers had started using the statements as mark-schemes. For a task addressing Strand A in Attainment Target 1, teachers might tick off 'distinguish between different kinds of historical change' (Level 5) or 'show an understanding that change and progress are not the same' (Level 6). Meanwhile, over in causation, we moved from 'identify different types of cause' (Level 5) to 'recognise causes can vary in importance' (Level 6) to 'show how causes of an event are connected' (Level 7).

If you use such summative statements as mark-schemes for ordinary work, you start to create paths to desperate history. You start to reward absence of thinking and absence of knowledge. The statements could be gamed with a formula. All you had to do was memorise four 'types' of change and slap them onto the content whenever a suspected change hove into view. Make links between causes? Easy. Go link crazy. Hey presto, Level 7.

Even when taken seriously, such statements could not reflect a meaningful hierarchy of difficulty. Linking or prioritising causes can be tough or easy for 8-year-olds... or for post-doctoral students. Difficulty depends on the range, complexity and specificity of content, as well as the question asked.

The disconnect between knowledge and disciplinary thinking had other strange consequences. Some teachers just taught the content, and then reached for an 'assessment' based on the levels! This was weird in two ways. First, how could pupils discuss causality without any focused practice in it? Second, how could this give assurance that pupils were gaining adequate knowledge?

The assessment framework began to affect teaching in bizarre ways. Some history teachers put 'statement of attainment' ladders on walls or even made pupils memorise them.

All-change in 1995! Lessons learned?

By 1993, it was clear that something must be done. Even those teachers not yet realising the above dangers complained about the workload of recording against all those criteria. Similar problems in other subjects led to a call to slim down the whole NC.

Fast forward to the 1995 NC – a very different beast. No more multi-stranded attainment targets with numerous tiny statements about little bits of skill or thinking. No more separation of disciplinary thinking from the substantive content. Now, at each level, all facets of historical accomplishment were blended in a single Level Description, a longish paragraph which could only be used in a 'best fit' way, at the end of a key stage.

This last was crucial. The new Level Descriptions were not to be used on single pieces of work. At policy level, even the civil servants were clear: a final performance should *not* be confused with the means of its nurture! Never again would formative assessment ape summative statements! Never again would *learning* be confused with *performance*!

For ordinary, routine assessment, from week to week, teachers were now to use whatever worked best to check that pupils really were building knowledge and thinking historically. They could use vocabulary tests, timelines, reading tasks, essays, diagrams... whatever teacher judgement decreed. Teachers would assess what they had taught.

Lest anyone made the dreadful mistake of using the summative levels for assessing ordinary progress, official guidance went to great lengths to stress that Level Descriptions were not even for *end-of-year* assessments; they were for *end-of-key-stage*. Pupils could now get better at history in focused, history-sensitive ways, building wide knowledge and tackling a growing richness of historical problems.

For a while, fruitful experimentation flourished.

History repeats itself.... twice!

If you've taught for twenty years, you'll know that this tale was repeated ...twice.

The first repeat began in the early 2000s. Experienced history teachers watched in horror as school leaders began to demand that Level Descriptions (often chopped into even tinier bits than the 1991 NC) be applied to individual pieces of work. In 2004, Ofsted's history lead criticised this practice. In the same year, Burnham and Brown (*TH 115*) launched their rallying cry of resistance:

Turning Level Descriptions into little stages presents history to pupils as colouring by numbers: 'If I include 4 good reasons, make links, use paragraphs and find two evidence examples, I'll be at Level 5.2X!!'.*

But by 2010, to resist was to hold back the sea. By then, even Ofsted expected pupils to know which level they were on! Summative assessments were being used for a purpose for which they were not designed.

The second liberation occurred in 2014 when Level Descriptions were abolished. But collective memory was short. History repeated itself, yet again. Into many schools in England came another levelling system – another skill hierarchy – the GCSE mark-scheme, and even GCSE grades. No policy change led this. No one mandated it.

But school leaders wanted what they *thought* was precise prediction.

This was all the problems of 1991 plus new ones. GCSE markschemes do not even purport to be a progression model. They are not a journey towards improving historical thinking or knowledge, nor even towards success in GCSE itself.

For the third time in twenty-five years, rigorous history assessment was forced underground.

History teachers to the rescue

A much more positive story runs alongside this. During brief periods of liberation and darker days of wonky systems, many history teachers never stopped exploring useful assessment. They never stopped debating how to define a gold standard in history and check on progress towards it.

And for now, at least, some wider contexts of inspection, policy, assessment theory and research are more favourable to sensible assessment. Well beyond the history classroom, a head of steam has built *against* the use of generic skill hierarchies to capture progress and *against* frequent summative assessments to show it. We might summarise these as follows:

- **Recent assessment theory**, such as that of Daniel Koretz, shows how student progress is undermined by teaching to the test rather than to a subject's full domain.¹
- **Cognitive science** emphasises how schemata in long-term memory fundamentally change how the brain perceives new material. We need multiple, connected reference points for speedy recognition of vocabulary and flexible thinking. Broad knowledge therefore matters. Funnelling assessment into the skill descriptors of final performance detracts from the build-up of background knowledge.
- **Ofsted's 2019 framework** reflects these shifts. The whole curriculum is what changes the pupil, so assess mastery of the curriculum. This is why Ofsted criticises the use of GCSE mark-schemes at Key Stage 3. The *curriculum* (as opposed to a skill hierarchy or final performance descriptor) is the progression model.²

The solutions that follow have not, however, arisen because of policy makers, nor because of Ofsted's change of heart, nor because of newly popularised research in cognitive science. The solutions predate these things. They have arisen from history teachers staring hard at what actually makes a difference for pupils, staring hard at history itself and refusing to accept practices that make nonsense data or terrible history.

And it will be down to history teachers to keep all this alive and renewed.

Key principles

1) Make the most of routine, informal assessment

Good formative assessment is ongoing, embedded and diagnostic. A major part of this is the informal assessment that forms a teacher's regular, ordinary feedback loop. Each time you listen to a pupil or read their work you are assessing what they have understood, what they have retained and how they can use it. Like all assessment, a measure of its usefulness is whether it helps you to improve your teaching and your curriculum.

Routine, informal assessment gives you instant feedback on your teaching. In their writing or speaking, a pupil might reveal a narrative confusion, fail to recall a suitable abstract term, use anachronism in a suggestion of causes or muddle the material conditions of two different settings.

This can influence your response on differing timescales:

- *You might fix it immediately* in your feedback, whether to whole class or individual, or in a re-worked lesson conclusion (see Worth *TH 173*). This could be as simple as requiring pupils to practise a phrase or promptly clarifying the implications of an historical state of affairs.

- *You might re-plan the next lesson.* Perhaps you decide that the remedy is to revisit an earlier story or to supply more period detail in the next. Perhaps you decide to add in a visual source that will obviate a misconception or challenge narrow, single-track historical reasoning.
- *You might re-think part of your curriculum* (perhaps with colleagues). Were our medieval African examples rich enough in that lesson sequence to obviate this confusion about evidence? Should our stories about the cult of the saints come much earlier, in order to ensure that medieval reaction to disease is understood? Do we need longer to prepare them to read a particular text?

Such responsive assessment is a natural part of good history teaching because all historical learning is dialogic and relational. It is dialogic because you are always trying to understand and to respond to your own pupils' developing mental models of the past and of how historians work. It is relational because you are always finding optimal moments for sharing your own fascination with the past, your own continuing study of it and your own connection with its ever-changing interpretation.

2) Vary your formative assessment: diagnose from all angles!

Knowing and doing history are multi-faceted. But if we regularly put *together* insights from different kinds of tasks, rather than relying on any one of them to do too many jobs, we will have a better understanding of a pupil's growing knowledge, vocabulary and thinking.

Some of these will be responsive and adaptive (such as those in (1) above). Some will be tasks which we've planned to embed in a lesson sequence. Some will be larger tasks, such as extended writing for answering an enquiry question or a formal test.

Such assessment moments need also to vary their focus. Fordham (copying Aristotle) called this a 'mixed constitution', by which he meant that no one approach can spotlight all aspects of how well pupils are mastering the overall history curriculum and being changed by it.³ For example, you might:

- Use speedy time-line tests to check pupils are secure in events just studied or where they fit into a chronology stretching across two years' work (use Carr and Counsell *TH 157*).
- Set up a pupil discussion involving use of a term. Burnham and Brown (*TH 115*) set short oral tasks in which they could hear immediately who was still unsure of the term 'imperialism', who was operating with a limited idea of the term and who was ready to be challenged by questioning its boundaries.
- Start each lesson by checking pupils remember essentials from previous lessons. Drawing inspiration from the 'Previously on...' technique of many TV series, Canning (*TH 179*) shares an imaginative and practical way to do this.
- Do quick knowledge tests to check that certain frameworks, vocabulary or dates are secure. Donaghy (*TH 157*) repeated the same tests at intervals.
- Have pupils develop and demonstrate their knowledge *and* thinking from a sequence of lessons by answering

the enquiry question that governed that sequence, in an essay or other extended task. For examples of this working in different types of disciplinary thinking, see other What's the Wisdom On features such as that on causation in *TH 175* or change/continuity in *TH 179*.

- Use a focused visual assessment. In *TH 130* Stanford describes a drawing task which captured his pupils' understanding of Renaissance ideas.

Remember that any such task doubles its value by *also* acting as the 'retrieval practice' which strengthens memory. It is therefore never *just* an assessment; it is also part of planned learning.

3) Isolate it ...

Each of the above involves either isolating things or integrating things. Assessment, like teaching, always involves deciding whether to isolate *components* or to integrate a *composite* of several components.

When Fordham developed the idea of the 'mixed constitution', this is what he meant. He drew inspiration from how music teachers assess progress in learning a musical instrument. This might range from playing scales, to aural tests which reveal theoretical knowledge and harmonic memory, to playing rehearsed pieces of various genres, to sight-reading with unseen music, to improvisation and composition. Too much of one, at the expense of the other, might distort the impression of progress.

History is just as complex. We need to check on underlying components and to look at the rich composites of open, interpretive tasks which require blends of recall and reasoning.

Isolating components might mean asking pupils to:

- Tackle simple recall questions, 'Who introduced...?' 'What do we call...?' When was...? A simple tot-up of marks can confirm who is and isn't secure.
- Answer multiple-choice questions designed to check for misconceptions (i.e. *not* just for straight recall). Stanford in *TH 168* illustrates smart multiple-choice questions such as: 'Would a medieval person think the following things?'
- Create or complete a timeline from memory, sequencing and/or dating events.
- Write just one paragraph to answer a tricky open question such as 'How powerful were medieval kings?' The pupils' range and comprehension of relevant historical knowledge will surface quite quickly, exposing gaps and faulty assumptions.

We can also shine a spotlight on just one feature through an extended task. As part of an enquiry examining medieval politics through stories of powerful women, Carr (*TH 184*) peers into pupils' extended writing to examine one thing – pupils' conceptions of medieval power and authority.

4) ...or integrate it

History teachers have traditionally drawn substantive and disciplinary knowledge together through an extended, end-of-sequence task in which pupils answer the enquiry question

that has driven the lesson sequence journey. As Brown and Burnham (*TH 157*) put it, 'These enquiry-based tasks ensure that assessment is integral to the teaching, bringing together the learning that has taken place rather than being bolted on at the end of a topic.'

Numerous articles in *Teaching History* illustrate this in detail. Mills (*TH 182*) has his pupils answer, 'Was there a mid-Tudor crisis in the reigns of Henry VIII, Edward VI and Mary I?' This question proceeds naturally from the previous eight lessons in which the idea of 'crisis' is repeatedly revisited and questioned through new content. The pupils' concluding essays allowed Mills to assess how nuanced his pupils' reflection on the concept of 'crisis' had become and how growing knowledge and focused problem-solving were achieving this.

Worth (*TH 184*) concluded a five-lesson enquiry with 'What do oral traditions reveal about the rise and fall of the Inka empire?' Her enquiry moved through different types of source – archaeological finds, landscape features, oral tradition and Spanish accounts. In the last lesson, Worth's pupils answered the enquiry question with two paragraphs, each on a different story from Inka oral tradition. Through these, Worth discerned how well her pupils were establishing and weighing evidence from diverse sources.

For more examples, see 'What's the wisdom on enquiry questions' in *TH 178*.⁴

5) Make the recording fit the assessment

Ongoing, embedded, diagnostic assessment does not necessarily need a numerical record in a mark book. But it might need a written note: perhaps some pointers for whole-class feedback, some reminders to tweak the next lesson or to give different support to Pupil X, some annotations to the curriculum to share with colleagues. What matters is that information from assessment is precise, meaningful (not buried behind numbers) and practically useful.

Burnham and Brown (*TH 115*) illustrate imaginative, simple and practical devices for mark books such as a sentence of qualitative comment or a simple coding system to highlight pupils who need specific additional work. As they note, this 'is extremely valuable data. It can be used by the individual teacher and by the whole department in assessing their effectiveness.' Burnham created a chart to show how pupils' use of the term 'imperialism' changed across the key stage, and how this improved her planning.

If you do use numbers, keep them close to the thing you are actually assessing, such as Donaghy's (*TH 157*) testing that fundamental knowledge is becoming more automatic. Keep any mark-schemes that are used on longer pieces of work both topic-specific and question-specific. See examples in Hammond (*TH 157*) and in Brown and Burnham (*TH 157*).

Mark-schemes always carry a risk, however, of fostering a formula. One way of avoiding this altogether is to use comparative judgement which is a form of norm-referencing against experienced teachers' instinctive disciplinary gold standard.⁵

6) Devise mixed summative assessments

The principle of a mixed constitution works in summative assessment too.⁶ If you have a twice-yearly exam, for example, then blend focused, component-specific questions (a timeline test, some short answers, multiple choice) and longer tasks such as a passage from scholarship for pupils to read and summarise (their recognition of vocabulary will indirectly assess their background knowledge) or a couple of essays requiring different sorts of disciplinary thinking (drawing on, but never exactly copying, the enquiry questions that they have tackled on, say, evidence or historical change).

Stanford (*TH 168*) tells the story of how one department went about designing end-of-year exams. They began by identifying the residue or 'takeaway' knowledge that they wanted pupils to have for each period, and then considered the substantive concepts and disciplinary thinking too. Stanford's final exam ranges from 'odd one out' questions to full essays. It results in a percentage.

Working out what matters in a summative assessment can be bewildering, so focus your minds by defining some 'takeaways' like Ian Dawson's on medieval history.⁷

Pitfalls

1. Don't *assume* that frequent low-stakes testing brings lasting or flexible knowledge. Dennis (*TH 164*) showed how securing knowledge in one format doesn't necessarily foster effective use in another.
2. Take care that your mode of assessment doesn't distort meaningful history. If we reduce all testing to chronologies and multiple choice, we lose the meaning that lies in history's natural forms of accounting – stories, debates and arguments – as well as opportunities for pupils to follow questions of interest to them. In *TH 164*, Luff reflects on how easily assessment slips away from such authenticity.
3. Above all, avoid leaning on a deficit model, where you use summative assessments to establish gaps and become driven by interventions. Summative assessment is a poor guide to *true* gaps.⁸ Stay focused on building security in the first place!

REFERENCES

- 1 Koretz, D. (2018) *Measuring Up: what educational testing really tells us*, Cambridge, Mass.: Harvard University Press. For a summary of Koretz see <https://daisychristodoulou.com/2014/01/why-teaching-to-the-test-is-so-bad/>
- 2 The term 'curriculum is the progression model' coined by Michael Fordham is explained here: <https://clioetcetera.com/2017/03/04/the-curriculum-as-progression-model/>
- 3 Fordham, M. (2013) 'O Brave New World without those levels in't: where now for assessment at Key Stage 3?' in *Teaching History 153 Supplement: Curriculum Evolution*.
- 4 The Historical Association has also recorded two webinars on using enquiry questions: www.history.org.uk/secondary/resource/9839/film-whats-the-wisdom-on-enquiry-questions-p
- 5 For information on comparative judgement, see <https://www.nomoremarking.com/>
- 6 Fordham (2013) *op.cit.*
- 7 Dawson, I. (2018) *Exploring and Teaching Medieval History in Schools*, Historical Association. Page 102 has Dawson's 'takeaway' statements on medieval history.
- 8 For why formative inferences from summative assessments are so problematical, see Christodoulou, D. (2016) *Making Good Progress? The future of assessment for learning*, Oxford: Oxford University Press.