Assessment knowledge

Over the last five or six years, there has been a significant change in the way that people think about the role of knowledge in the school curriculum, due in no small part to the works of E.D. Hirsch becoming better-known in England. Hirsch's book *Cultural Literacy*, published in 1988, summarised much of the research in cognitive psychology which shows that knowledge is vitally important for thinking, learning, and problem-solving. For various reasons, his ideas have become well-known in England over the last few years, and for those of us in favour of a knowledge-rich curriculum, these last few years have been heartening: whereas once, mention of the word knowledge led to evidence-free stereotypes about backward-looking Gradgrindian taskmasters, now it is possible to have a much fairer debate with reference to the kind of evidence that Hirsch has done so much to publicise.

However, although there are now exciting debates taking place about the curriculum, assessment tends to get less public attention. Newspapers would rather argue about whether Florence Nightingale or Mary Seacole should be on the curriculum than about the arcane details of the Angoff standard-setting method. And yet, because of the exam-focussed reforms of the past few decades, assessment is often the driver of curriculum. In many schools, the ring binder or pdf which contains the national curriculum will be barely touched. The exam specification, by contrast, will be pored over as though it is holy writ. Changes to assessment have a significant impact on how the curriculum is implemented, and interestingly, Hirsch himself has written extensively about assessment too. In the rest of this essay, I will outline three assessment issues which I think are particularly important, and suggest some implications for policymakers.

Authentic assessments

Authentic assessments are those which aim to represent more accurately the kinds of problems a pupil might face in the real world. So, for example, instead of a science question which asks pupils to apply the speed-distance formula, or a language question which asks pupils to identify a verb or a noun, an authentic assessment will place these kinds of problems within a more real-world context, such as a creating a brochure to help people decide how to pick a fast car, or an essay about the impact that language has on the reader. Such tasks may involve groupwork and different kinds of activities: the assessment expert Daniel Koretz gives an example of an assessment designed to test pupils' understanding of density which required them to work in groups to construct an aluminium boat out of foil (2008, p.222).

On the surface, these kinds of authentic assessments seem far fairer, because they test the types of things we really care about. However, they have many technical flaws. Precisely because they are so authentic, pupils can respond to them in a number of different ways, which makes reliable marking very hard. Tasks such as the one Koretz mentions also introduce irrelevant elements: what if a pupil understands the concept of density, but struggles to make a boat out of foil? Whilst such tasks have been designed to reward creativity, paradoxically, they can actually end up stifling it: in an attempt to make the marking of such tasks reliable, they are often accompanied by extensive rubrics which define a 'correct' method of solving the problem (Wiliam 1994, p.54). Pupils who respond in a more ingenious way may receive no marks at all. Such was the fate of many of the coursework tasks on the old Science GCSE: acceptably authentic answers to these could be found on the internet.

The alternative to such assessments is more structured items, such as short answer and even multiple-choice questions. Multiple-choice questions in particular have a terrible reputation in the UK, with progressives decrying them as soulless, and traditionalists as 'gimmicky' and easy to guess (Wiliam 2014, p.55)**.** However, they also have an extensive amount of evidence on their side. Contrary to received wisdom, they are capable of testing higher-order skills: in the US, the GMAT determines entry into some of the most prestigious academic institutions in the US, and it is composed largely of multiple-choice questions. In the recent past, many top universities required sixth-form students to pass the 'Use of English' exam: part of the exam involved reading a passage of modern English and answering some fairly challenging multiple-choice questions on it. In *The Schools We Need and Why We Don't Have Them*, Hirsch reviews the literature on authentic writing tasks and multiple-choice questions, and concludes that when assessing writing, the ideal balance would be an exam composed of two parts multiple-choice, and one part writing task (Hirsch 1986, p.187). This mix of tasks delivers a high level of reliability, as well as retaining an authentic element. Nor are multiple-choice questions only of use in national exams: as Dylan Wiliam has argued, multiple-choice questions can be very powerful when used for classroom formative assessment, because the existence of several wrong options allows the teacher to identify who has grasped a new concept, and who is still labouring under a common misconception (Wiliam 2014, p.55). In short, we could all benefit from moving away from our prejudice against multiple-choice items, and using them to improve both exam reliability and classroom assessment.

Teacher assessments

Similarly, teacher assessments seem, on the surface, to be fairer and more valid than exams. Exams can only test what pupils know in a narrow two or three hour window, when a pupil's performance might be hindered by illness or a disturbance at home. The teacher has knowledge of the pupil that spans more than just those two or three hours, and is therefore better placed to be able to give a fairer assessment of the pupil's abilities. There is some truth to this, in that variable performance on the day is one of the main sources of exam unreliability. However, teacher assessment has significant flaws of its own. It is extremely hard to ensure that all teachers are applying the same standards in the same way. One review of the literature speaks of the 'depressing fallibility' of teachers' judgments (Sadler 1987, p.194). There is also evidence to show that teacher assessment is unconsciously biased against certain groups: disadvantaged pupils, pupils with SEN and pupils from some ethnic minorities actually do better on tests than on teacher assessments (e.g. Shorrocks 1993, Harlen 2004, Campbell 2015). This is a well-documented finding which is relatively little-known: indeed, it is not uncommon to find educationalists who assume the complete opposite, and argue that one of the advantages of teacher assessment is that it *benefits* such underprivileged groups (e.g. Bousted 2013, Emery 2013). Finally, teacher assessment is often extremely onerous, imposing a significant logistical and bureaucratic burden on teachers.

The above arguments may sound excessively critical of teachers. This is not the case: if teachers are bad at making such judgments, it is not because they are teachers, but because they are human. A growing body of research shows some of the difficulties everyone has with making certain complex judgments and decisions, and the short cuts we resort to when the mental strain becomes too great. Indeed, it is plausible to speculate that the reason why teacher assessment is biased is because it is so burdensome: when we are faced with difficult cognitive challenges we often default to stereotypes (Kahneman 2011).

Teacher assessment has already been reduced at GCSE because of some of the reasons outlined above. Currently, it is still used in the national assessments at the end of Key Stage 1, when pupils are 7, and there is a strong case for the government to consider whether these assessments are serving the best interests of both teachers and pupils. Possible alternatives are formal tests, or abolishing the assessments entirely, both of which would be controversial. But given the flaws outlined above, some form of reform is surely worth the controversy.

Criterion-referenced assessments

Criterion-referenced assessments are those where pupils are judged according to whether or not they have met a certain criterion – for example, whether they are able to use percentages, or whether they are able to punctuate sentences correctly. Again, on the surface this seems fair, as it means pupils are held up to an objective external standard. It certainly seems fairer than one of its main alternatives, norm-referenced assessments, where pupils are instead judged with reference to how other pupils did on the same assessment. However, in practice, the apparent simplicity of criterion-referencing is fraught with difficulty. What does it mean to say that a pupil can use percentages? That they can calculate 50% of 200? 67% of 5834? Or that they can solve a word problem involving the percentage profit on a series of goods that are all sold at different prices? Criteria can be interpreted in many different ways. Even simple changes in the structure and wording of a question result in vastly different numbers of pupils answering it correctly. More pupils will answer the sum 11+3 correctly than will answer 3+11. 90% of pupils can work out that 5/7 is larger than 3/7, but only 15% can identify that 5/7 is larger than 5/9 (Hart 1981). And these examples are from maths, a subject where criteria can be relatively precise. As Hirsch has shown, the problems with criterion-referencing are even more pronounced in English, where the criteria are often as nebulous as ‘can draw inferences such as conclusions or generalisations’ (Hirsch 2006, p.99).

As a result of the difficulties with criterion-referencing, public exams in England have never been fully criterion-referenced: exam boards and Ofqual quite rightly do not solely depend on criteria to set standards. But whilst few policy changes are necessary in this area, there are other problems. The recent Carter Review of Initial Teacher Training (2015) found that training in assessment was particularly weak, and that many important assessment concepts, including norm and criterion-referencing, were simply not being taught. This is of course problematic in and of itself, because such concepts matter. However, it is also problematic because one of the current government’s major policy aims is to create a school-led education system. As part of this, schools have been given the responsibility for designing a replacement for national curriculum levels, a criterion-referenced form of assessment which has been abolished by the government. But if initial teacher training does not equip teachers with key assessment concepts, then schools will struggle to design the reforms demanded of them. And indeed, the early signs are that a number of school’s replacements for national curriculum levels are simply rehashing a criterion-based approach to assessment (DfE 2014). There are fewer signs of schools using genuinely innovative replacements for levels, such as the No More Marking system of comparative judgment, which allows teachers to stop using criteria altogether.

Conclusion

A central theme of Hirsch's work is about the importance of ideas. More than one of his books features this famous Keynes quotation.

> The ideas of economists and political philosophers, both when they are right and when they are wrong, are more powerful than is commonly understood. Indeed the world is ruled by little else. Practical men, who believe themselves to be quite exempt from any intellectual influence, are usually the slaves of some defunct economist.

Ideas and research about the importance of a knowledge-based curriculum have made themselves felt both in the US and in the UK. However, new ideas about assessment are much less well-known, and there are many practical policymakers and educationalists in thrall to defunct ideas about soulless multiple-choice questions and straightforward criterion-referencing. Whilst there are specific policy changes the government could make which would help to improve assessment, the most significant change that could happen would be in the terms of the debate – and, as with the curriculum, E.D. Hirsch's work could be the catalyst for a change in how we think.

References

Bousted, Mary. (2013). Qtd in 'Coursework abolished as Gove brings GCSE shake-up', http://sqmagazine.co.uk/2013/06/coursework-abolished-as-gove-brings-gcse-shake-up/

Campbell, T. (2013). Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment. *Journal of Social Policy*, 1-31.

Carter, Andrew. (2015). *Carter review of initial teacher training.*

Department for Education. (2014). 'Schools win funds to develop and share new ways of assessing pupils'. https://www.gov.uk/government/news/schools-win-funds-to-develop-and-share-new-ways-of-assessing-pupils

Emery, Hilary. (2013). Qtd in 'GCSE proposals could disadvantage vulnerable pupils, fear experts' <http://www.cypnow.co.uk/cyp/news/1077476/gcse-reforms-disadvantage-vulnerable-pupils-fear-experts>

Harlen, W. (2004). *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Hart, K.M. (Ed.). (1981). *Children's understanding of mathematics*: 11-16. London, UK: John Murray.

Hirsch Jr, E. D. (1996). *The schools we need: And why we don't have them*. Anchor.

Hirsch Jr, E. D., Kett, J. F., & Trefil, J. S. (1988). *Cultural literacy: What every American needs to know*. Vintage.

Hirsch, E. D. (2006). *The knowledge deficit: Closing the shocking education gap for American Children*.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Keynes, J. M. (1964). *The General Theory of Employment, Interest, and Money.* New York: Harcourt, Brace & World, 1964.

Koretz, D. M. (2008). *Measuring up*. Harvard University Press.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*(2), 191-209.

Shorrocks D, Daniels S, Staintone R, Ring, K (1993). *Testing and Assessing 6 and seven year-olds. The Evaluation of the 1992 Key Stage 1 National Curriculum Assessment*. UK: National Union of Teachers and Leeds University School of Education.

Wiliam, D. (1994). Assessing authentic tasks: alternatives to mark-schemes. *Nordic Studies in Mathematics Education*, *2*(1), 48-68.

Wiliam, D. (2014). *Principled Assessment Design*. Specialist Schools and Academy Trust.