

'We have to remember that what we observe is not nature herself, but nature exposed to our method of questioning.'

Werner Heisenberg⁹⁰

This statement by Heisenberg relates to his work in the field of quantum physics which led to the development of his famous Uncertainty Principle in the 1920s and '30s. It is not possible to determine both the momentum and position of a particle with absolute certainty at any given time; this uncertainty is an inherent feature of the quantum universe and goes beyond being merely an observer effect whereby the process of measurement interferes with the process in hand.

However, the spirit of the quotation resonates loudly with so much of our thinking about assessment in education; just replace 'nature' with 'learning'. In recent years I have explored this issue in numerous blogs with titles such as 'The Data Delusion' and 'The Assessment Uncertainty Principle' – ideas I'll return to in this chapter.

So many teachers around the world are operating in conditions straight from the plantation where the value given to the assessment of student outcomes in the form of measures, grades and qualifications has been elevated to such a great height, with such high-stakes consequences for success or failure, that the fundamental principles of assessment have been distorted beyond all meaning. If we're going to create conditions for everyone to thrive in the Learning Rainforest, we need to put that right.

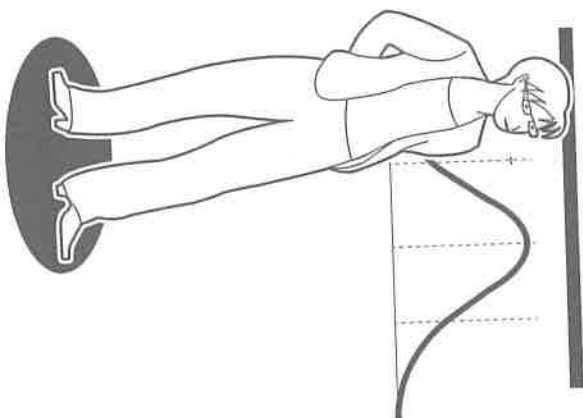
At a basic level, assessment is simply any process by which we can gauge the quality of students' learning via the proxy of some form of performance, using a set of criteria that determine the range of what we regard as standards relevant to the learning in hand. We then decide to use the assessment information to contribute to the learning process itself – feedback – or we use it to communicate our findings to anyone with an interest. We might well do both. As it is often complicated and cumbersome to communicate the details of what someone has achieved, we have invented a host of numerical and alphabetical systems to try to serve as scales – to keep things simple. And this is where the problems begin.

The current prevailing assessment paradigm is dominated by what one might call 'macro summative attainment tracking'. This has reigned supreme in England for the last 15 years or possibly more and is common around the

HOW DOES ASSESSMENT WORK?

CHAPTER 5

1 WE NEED A REAL PARADIGM SHIFT AWAY FROM MACRO SUMMATIVE DATA TRACKING TOWARDS AUTHENTIC FORMATIVE ASSESSMENT.



2 STANDARDS ARE ALWAYS DETERMINED BY REFERENCE TO THE BELL CURVE WHETHER WE LIKE IT OR NOT.

3 "DEFINING THE BUTTERFLIES" IS THE GOAL IN EVERY CONTEXT. WHAT DOES EXCELLENCE LOOK LIKE AND HOW DO WE COMMUNICATE THAT TO STUDENTS?

4 LOW STAKES HIGH-FREQUENCY TESTING IS A POWERFUL FORM OF FORMATIVE ASSESSMENT.

5 AUTHENTIC ASSESSMENT HAS DIFFERENT FORMS IN DIFFERENT SUBJECT DISCIPLINES. IMPOSING A GENERAL GRADE SYSTEM LOSSES THE INFORMATION THAT SUPPORTS ACTUAL IMPROVEMENT.

6 RESPONSIVE TEACHING, REQUIRES USING STRATEGIES THAT ELICIT INFORMATION ABOUT WHAT STUDENTS HAVE LEARNED AND GIVING THEM FEEDBACK THAT MOVES THEM FORWARD.

⁹⁰ Heisenberg, W. (1958) *Physics and Philosophy: The Revolution in Modern Science*.

world. The focus has been on trying to represent students' attainment across all disciplines in terms of generic ladders of grades and levels, occasionally supported by rubrics and descriptors, so that attainment and progress can be tracked. The value given to data tracking has been a driving force in educational management thinking; the idea that summative macro data is a requirement for driving up standards is widely held.

Elements of this paradigm include the following:

- Confusing bell-curve ranking with absolute standards and presenting one as the other alongside the false assertion that grades, by themselves, communicate information about learning.
- The illusion of 'progress' as something that can be measured via reference to bell-curve grades or data points in general; the very idea that 'progress' in learning has a measurable size, ludicrous as it is, is still widely held onto.
- The target-grade culture: the idea that, by setting systematic attainment targets to aim at, in the language of summative grades, students will learn more because students and their teachers will 'do better'.
- The attempt to use summative tests formatively. This is very common with students sitting endless mock tests and using 'exam-style questions' prematurely. This is akin to getting students to play a whole piano piece over and over before they've learned their scales or endlessly practising match situations in sport without drilling on basic skills.
- Schools' cultures around book marking where how the marking looks (to parents, inspectors, school leaders) has been more important than what it achieves with an appallingly low impact to workload ratio. Most marking is wasted and yet is often the cause of major teacher overload.
- Massive centralised data-tracking machinery where learning is represented in data form: schools across the world have got sprawling colour-coded spreadsheets where numbers and grades are entered for the purposes of tracking students' performance. Tracking may have its purposes in terms of describing where students are on a crudely defined flight path, but unless this feeds back into different actions in the classroom you can stare at the numbers and grades all day long and no child learns anything more as a result.

The trouble with macro data-tracking is that the journey from the spreadsheet to the classroom is circuitous at best, and normally doesn't happen at all. All

we are doing is generating data that might tell a story about relative student performance – it might, at best, give a picture of where students are – but it does nothing to take them further. This is because the grades themselves do not actually tell us much at all about what students know or can do at a level that is actionable.

If we are going to break out of this paradigm, to engineer a paradigm shift, we need to build a better one that is more fully focused on students' learning, building the knowledge structure that is so necessary in creating the lush canopy in our Learning Rainforest. To do this we need to agree on some key principles and explore some misconceptions about assessment, adopting a mentality that mirrors a shift from Plantation Thinking to Rainforest Thinking.

5.1 Assessment, standards and the bell curve

There are lots of well-worn clichés in educational discourse around assessment and testing that are problematic; often challenged for the wrong reasons. It's worth exploring them.

'Weighing the pig doesn't make it fatter.'

If we are talking about the assessment data we collect for tracking purposes then yes, this analogy works – but only if the data does not feed into any programme that might adjust the conditions that determine what the pig eats and how it lives. As with the literal pig-weighing scenario, educational assessment data has no intrinsic value in relation to improving the quality or depth of student learning; that would only be true if we use it to inform our understanding of how the teaching and learning process might need to change. So, of course, if we do use assessment data formatively as part of a feedback loop, then the pig might well get fatter because we weighed it and learned something. The question – which we'll return to – is: what might that kind of data or information look like?

However, as we now know from cognitive science – for example from Bjork's work on desirable difficulties referenced in Chapter 4 – the process of testing can in fact strengthen retrieval and improve our retention of knowledge in our long-term memory. In that context, the act of weighing our pig is actually going to make it fatter; regardless of the measures we use, testing that supports retrieval of knowledge we've already encountered can support deeper learning. Too often, people talk in damning terms about testing when it is really the value placed on test outcomes and the way they are used that are the problem.

If tests are low-stakes, the very act of being tested has a positive learning effect. We need to bear this in mind. If we are going to be serious about building a

strong knowledge structure for all the specimens in the Learning Rainforest, we need to embrace testing as a positive tool in our armoury. Tests are good! This is especially true if the curriculum being tested is quite tightly defined so that the test scores are raw marks that reveal some specific information about what the teacher might need to work on in order to push his or her students further.

One of my provocations to schools leaders is to consider the scale of the data machinery that they create. A favourite Heath Robinson drawing depicts a complex arrangement of cogs, levers, pulleys and string that supposedly facilitate the process of peeling potatoes one by one. The alternative? You could just use a peeler. I enjoyed the response to this 'thought experiment' that I once posed on Twitter. Try it:

Your entire assessment database is wiped (Oops!). How long before this has an impact on a single student?

Or try this: Ask your colleague teachers to write down the names of their students who they feel are excelling and those who need some special attention to get them back on track. Compare this with the data. How did they do? When you factor in assessment error and the time taken to produce the spreadsheet, teachers' knowledge of their students probably stacks up pretty well. For sure it will be more nuanced than the crude data representations. So, let's be sure not to equate 'assessment' with 'data'. That's missing by a mile.

'Raising the Bar.'

With the classic high-jump metaphor the suggestion is that, by making assessments more challenging, we will also be raising standards of attainment automatically. As with the real high-jump, this could work in situations where our learners were always capable of achieving higher standards but expectations were too low – or, if, by raising expectations, learners become more motivated and consequently strive to succeed at a higher level. I've known plenty of situations where that is true – and it echoes with the Pygmalion effect mentioned in Chapter 4.

However, if a student is already struggling to jump over the bar, despite a high level of motivation and effort, then raising it is unlikely to lead to greater success. Harder exams don't, in themselves, raise standards. They tend to have the effect of stretching out the distribution of scores but they don't make anyone a better learner unless other effects come into play. You need to be very sure that the low expectations effect is the key reason for underperformance before making the learning more difficult, especially as, as discussed in relation to

mindsets, a high initial success rate is needed to build confidence and generate a growth mindset in any given situation.

'The exam factory'; 'teaching to the test.'

These are often part of the well-rehearsed rhetoric from those who protest against the dominance of examination outcomes over our school system. As we discussed in Chapters 2 and 3, there is a lot more to a 'great education' than accumulating examinable knowledge. If a school is an 'exam factory', the suggestion is that exam results matter more than they should, that decisions are taken that might yield stronger examination results but might not necessarily provide students with the full 'philosopher kid' experience.

At a more granular level, within each class, 'teaching to the test' is almost always offered as a critique of the effect high-stakes testing has on the enacted curriculum. Instead of a test sampling a wider curriculum, the curriculum is confined to overlap closely with the possible content of any test. It's clearly the wrong way around.

However, the danger of both of these high frequency edu-memes is that they overly diminish the positive or at least necessary role that assessments, examinations and testing play in our education system. Do we want to have well-respected qualifications that indicate a level of mastery in a particular field? Do we think awarding qualifications should be based on examinations and other formal assessments? Are they a reasonable means of gaining a measure of students' learning – their capacity to deliver in Briceño's Performance Zone? Do we want our public examinations to be well-regarded as a valid, reliable indicator that certain standards have been met, carrying a degree of public confidence, recognition and prestige? Do we think it is fair for learners to be able to be set tests that are reasonably predictable within a range, so that they can prepare adequately? Is that not fair?

As ever, there is a degree of balance required, especially in the context of an education system where socio-economic disadvantage plays such a big role. We can't allow our justified concerns about the excessive status of exam outcomes to perpetuate inequalities by approaching exams for disadvantaged children with any less rigour than the more privileged school up the road. As Headteacher John Tomsett says, sometimes the best pastoral care we can give a disadvantaged child is a good set of exam results⁹¹. In the real world they live in, this is their ticket to opportunity. Which leads me to the last of these statements:

91. Tomsett, J. (2014) johntomsett.com/2014/07/20/this-much-i-know-about-what-year-7-pupils-mums-really-worry-about-and-why-your-keepy-appy-skills-really-matter/

'We value what we measure when we should measure what we value.'

In terms of public examinations, this is clearly true. We should reflect on the value we place on examination results relative to all the other elements that constitute 'a good education'. Exams don't capture the extent to which young people contribute to their communities; their capacity for moral leadership; their commitment to artistic and sporting excellence; the empathy and support they give to people living in poverty across the world or their willingness to tackle prejudice.

Exams don't provide a measure of how effectively someone works in a team; their ability to take initiative or to respond calmly in a crisis; their ability to undertake a long-term in-depth study; their skills in dealing with people from different backgrounds or their powers of persuasion or to be thoughtful, kind and tolerant. Obviously enough, exams only measure what can be measured in exams.

Meanwhile, as a society, we place considerable value on all the other attributes. In the ideal world, our diverse, thriving rainforest education would encompass a system that takes greater account of this bigger picture and allows young people to show more clearly what strengths they have and the full extent of their personal achievements. I'm currently working with people from across the education sector in England to devise a qualifications framework that gives greater weight to these other dimensions – The National Baccalaureate for England⁹². Taking inspiration from the International Baccalaureate, our model places academic learning at the core whilst also giving credit for aspects of learning that are not included in exams but that employers and colleges value. It's a grassroots initiative which means it takes time to build – but I am convinced that its time will come.

In the meantime, it is vital that we continue to remind ourselves that, however important exams and qualifications are, they only tell a part of the story of learning and achievement for any student and for any school. This does not mean we should start to engineer arbitrary measures for student qualities and characteristics – to turn qualitative evaluations into data. The folly – the horror! – of trying to measure children's character as they move up the Resilience Scale or the Collaborative Learning Index. If you ever find yourself down that blind alley, turn around and run. It's a dead end. We don't need to turn everything into data in order to give it value.

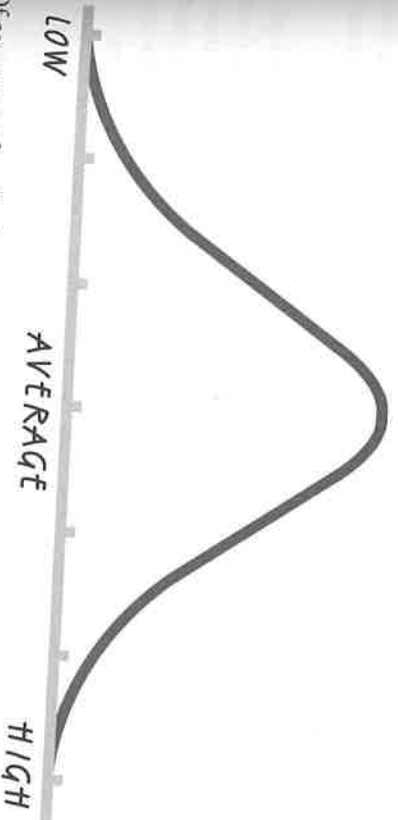
92. National Baccalaureate Trust www.natbaccrust.org

Absolute standards vs relative standards

At the heart of our problematic relationship with assessment lie two contradictions. One is that tension we've just explored between the assessed curriculum and the broad curriculum we value. We might all recognise that formal assessment is an important element in the big picture and accept that only part of our curriculum can be assessed in a reliable manner; at the same time we are determined to give value to the learning experience in all its diverse glory, going far beyond what can be measured in tests. It's not an easy balance to strike in a climate of hyper-accountability where data is king.

The other key tension is our desire to give all of our learners the opportunity to succeed, whilst also holding a strong desire to establish standards that must be met at various points along every student's educational pathway. Here we need to clarify what we mean by standards; we need to face the reality that absolute standards are very difficult to isolate from a deep-seated devotion to a bell-curve distribution where, in essence, learners are in competition with each other.

The bell curve



Of course we should also acknowledge the limitations a bell-curve philosophy imposes on us. 'Shock!! HALF of all students are BELOW AVERAGE!!'. 'ONLY 2% of students are good enough to get the top 2% grades!!'. 'ONLY newspaper editors don't seem to understand that these data scandals are a joke. Perversely, if things go too well, we're still in trouble. There is a fairly profound schizophrenia that leads to people jumping up and down about standards slipping because too many people gain the highest grades. This is 'proof' that standards are falling. Grade inflation undermines public confidence in

standards. It's a deeply held principle that accolades of any kind are cheapened if they appear to be freely attainable.

However, before we get our tinfoil hats out looking for injustice, let's examine our own sense of what we mean by standards. What about marking an essay? We can all acknowledge that marking extended writing is a messy business. We can give an impression mark using some criteria; we can count up some definable features – we can give a score out of 30 referencing some exemplars. All of this is complex and, in truth, there is no one correct mark for something as complex as an essay or any form of composition.

Try getting two or three people to agree the mark or grade on ten pieces of work. Then scale that up to 2000 markers across the country and see the scale of the issue. What is a boundary for a B-grade? How do you measure it? How consistent are we from one person to another or even as individuals, from one day to the next? It's hard just to agree with yourself over the space of a few hours of marking.

On a comprehension paper, if the marks vary from one year to another, how do we know whether it was a hard paper or a less successful cohort of students? We need reference points and these generally arise from the cohort. It is a safer bet that each cohort of students has a similar ability profile year on year than that the tests we set each year are of the same standard – especially on a national scale.

Let's look at something that should be purely objective and absolute. The high-jump. Imagine 1000 children being asked to do the high-jump, giving them a few attempts to get their personal best. There is no limit imposed on them. Imagine Jane goes home and tells Dad, 'I jumped 120 cm today'. He's delighted of course...at first. But then, after a moment, he asks, 'Was that good? How high did everyone else jump?'

It's the obvious question. Without any reference point, 120 cm means little. It could be very high, average or very low. Dad needs to know what the background profile is in order to gauge his child's achievement. 190 cm is a low exceptional jump because of how it compares to everyone else. 50 cm is a low standard for high jump – but only because almost anyone could jump this high. Of course we can talk about progress, about personal bests, personal triumphs, the impact of disability and so on, but every child will also want to know where they stand along the line before their personal achievement makes sense.

The same is true of piano exams. In various instrumental grading systems such as ABRSM⁹³, grades 1-8 are chosen at intervals of difficulty that span the typical range of human ability at playing the piano. It would be daft to set Grade 1 so high, no-one could pass, or so low that anyone could pass. It represents a fair challenge – a standard – for beginners. This is uncontroversial. Even within the cohort of Grade 1 entrants, there is a range. A bell curve. Some fail, some pass, some get Merit and some get Distinction. These measures are defined with reference to a background spread of ability within what is possible given the pieces and scales (the curriculum) set for Grade 1. There is a good sense of piano exams being absolute standards, but every aspect has an origin in relative standards. The success and stability of ABRSM exams lies in the extent and quality of the moderation processes that the examiners are required to undertake.

We might like the idea of fixed benchmarks that anyone can meet but, in practice, do we really?

- Elif can spell 'accommodation'.
- Sam can multiply any three digit number by any other.
- Ruby can explain the formation of Durdle Door in Dorset.
- Eleanor can write a coherent paragraph analysing Othello's character flaws.
- Joshua can work out the nth term of a linear sequence.
- Daisy can state which of two compounds is formed via ionic bonding.

Naturally enough, it turns out that some perfectly sound explanations of Durdle Door or analyses of Othello's flaws are much more sophisticated than others; the seemingly straightforward assertion that someone 'can do' something is often far too simplistic. It might be true that Samir can define 'connotation'; he might be able to think of a connotation for the word 'red' used in a poem. But, given various alternatives, could he determine which connotation from a set of plausible alternatives is the most appropriate in the context of the imagery intended by the poet?

Whilst 'can do' statements appear to be absolute, if they are used to set standards, they are no use unless we know how difficult they are. So, we ask thousands of children questions that tease out whether they can do these things and see what we get. If lots of children can do something, we say it is easier

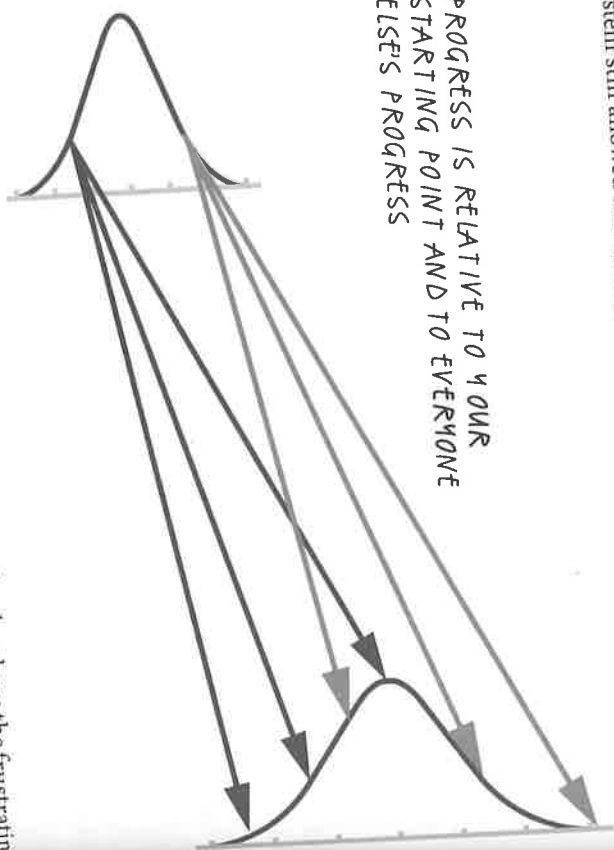
⁹³ Associated Board of the Royal Schools of Music www.abrsm.org

than something only a few children can do. Absolute standards only fully gain meaning in reference to a cohort; that's an inherent feature of our sense of 'difficulty'.

The paradox we face is that we want all children to be able to succeed; we're uncomfortable with the idea that there are only ever a limited number of positions on the achievement scale that we might consider 'excellent' or even a pass. And yet, at the same time, a *de facto* competitive, comparative process underpins nearly all our attempts to define standards.

The solutions might lie in giving more value to every position along the scale. When my son took piano exams, in passing Grade 1 and all the grades up to Grade 5 – by which time he'd had enough! – he did not feel that he failed Grade 8. He started out at the bottom end of the bell curve and made his way along at his own pace. The structure of assessment built confidence at each step whilst remaining challenging continually. He always knew where he stood but the system still allowed him to feel that he was succeeding.

PROGRESS IS RELATIVE TO YOUR STARTING POINT AND TO EVERYONE ELSE'S PROGRESS



If only our public examinations were like this. In England we have the frustrating situation where, on a scale of 1-9 in our 16+ GCSE exams, 9 is reserved only for the top 2% or so; only grades 4-9 are deemed by the government to be 'good grades' so that grades 1-3, by extension, are regarded as 'bad grades'. Approximately 30% of students must get bad grades by definition; it's locked

into the grade-award mechanism, a hard-wired technicality. The assumption is that a student with a Grade 1 has simply failed; there is no positive value at all given to the elements of knowledge that they do have; all we know is that they've performed less well compared to 90%+ of their cohort.

Another way to harness the bell curve is to use it to focus on progress. Given that some students will continually struggle to match the standards of their peers, we can at least look to see how much progress they make. Quantifying this is problematic but, using comparative measures of various kinds, we can establish how well students are doing compared to their relative starting points. It could be that a low attainer makes excellent progress compared to their peers – as illustrated in the diagram. Similarly, we can identify whether a higher attainer continues to make progress or whether their attainment level is masking a drop-off in their progress. In terms of identifying success stories and students causing concern, relative progress is as secure a concept as fixed benchmarks, given how difficult benchmarks are to pin down.

The data delusion and the assessment uncertainty principle

As well as the constraints of the bell curve, I think teachers and leaders should give greater weight to the inherent limitations of awarding marks and grades to pieces of learning. Grading is such a massively flawed process but it still holds sway across the world, protecting an illusion of certainty and precision that isn't actually there. It still astonishes me that across the US, the Grade Point Average system retains such high status when there is so little benchmarking. There are no robust mechanisms at play to make sure that a GPA of 3.3 for a student from School A has the same value in terms of standards of learning as the same score for a student from School B.

Grading works as long as we are clear that they are essentially impressionistic; broad distribution markers with fuzzy overlapping edges. If Elif, James, Hassan and Louisa all write essays about *Of Mice and Men*, a range of markers might not agree the marks for each one because it depends how each marker interprets the criteria. However, they could probably agree on the rank order for the four essays. With the help of some graded exemplars, they might even agree where the essays fall within broad standards that could be labelled A, B, C.

But, let's look at a grade boundary scenario. What if Elif's score in a set of assessments made of several questions is 31/60 and Louisa's is 32/60, where 32 is the pass/fail boundary? The assumption that Louisa has 'more knowledge' or is 'better at writing' than Elif is highly dubious. It really depends on which questions they got right and which fairly arbitrary criteria are being used to award marks. Even if they had the same score, they could have a very different

range of knowledge about *Of Mice and Men* and the process of writing in the appropriate style. The score itself tells nothing about what each of them has learned. This is true of all data outcomes from assessments; as we move from marks to total scores to grades, we're getting further and further away from knowing any single detail about what students know.

Importantly, it's also a fact that, with a different combination of equally plausible mark weightings for different parts of the assessment, Elif and Louisa could reverse their scores. And yet that cliff-edge – where 32 is a Pass and 31 is a Fail – could have real-life consequences for those students. There's an element of arbitrariness in the mix that doesn't have enough recognition. I'm especially sceptical about A-/B+ grading. A former colleague told me that his university tutor would award grades such as B⁹⁰ as distinct from B⁹¹. The tutor was adamant that she could apply those grades consistently for a history essay. The truth is that it's a delusion to believe that anyone's judgement of grades has anything close to that level of precision and reliability.

I've always loved this story: A colleague at KEGS was challenged by a parent at parents' evening: *Why is this essay only worth a B?* The teacher looked at it and then proceeded to explain how he'd applied the assessment criteria, leading to a B being the best fit. The parent wasn't happy: *'I still think it's worth an A.'* There was a heated debate. The parent felt the teacher wasn't giving his son the credit he deserved. After a while, the teacher, very wise but slightly weary, said 'OK'. He took the essay back, got out his pen and wrote 'A' on the top before handing it to the parent. 'There you are'. Had this changed the quality of the essay? No. Did it change the true value of the essay? No. Did it change the teacher's judgement of the actual standard? No. It was a Pyrrhic victory for the parent – who knew this all too well.

I love this because it shows how precarious and shallow the whole notion of grading can become. What is an A? How can we be sure? Who is it for? Most importantly, we should be asking – what does a grade or score tell us about what students know and can do so that we can help them to improve? Imagine coaching a tennis player to improve their technique. How much use would it be to tell them that their serve deserved a B+?

5.2 Formative and summative assessment

In the rainforest, growth actually happens at a cellular level, not at the level of trunks, branches and leaves. That's just what we can see; what we can measure as an overall effect. The same is true of learning; it only happens in the detail of what students know and understand and the things they can do. In order to get past the limiting bell-curve grading issues, a part of the solution is to focus

less on the graded representations and more on the actual elements of what is being learned.

There are three sets of ideas that I want to explore here:

- Daisy Christodoulou's ideas about separating formative and summative modes of assessment to suit their different purposes
- The idea of formative assessment as expressed by Dylan William and others – what he now calls 'responsive teaching' – and the role of feedback within that
- The idea of authentic assessment, which prioritises meaningful measurement appropriate to subject disciplines over the need to provide comparable grading measures

Making Good Progress

In her excellent book *Making Good Progress*⁹⁴, Daisy Christodoulou performs a thorough dissection of prevailing assessment practice. Drawing extensively on ideas from cognitive science and core knowledge advocates such as E D Hirsch, she makes a compelling case that we are often using the wrong forms of assessment for the purposes we seek. This has some significant consequences for the breadth and depth of our curriculum, our understanding of standards and the tools we use to raise them. She concludes by setting out exactly what the features of effective formative and summative assessment should be.

Formative assessment should be geared towards 'identifying consequences'. This kind of assessment is all about providing information that can directly feed back into the teaching process. To be effective, assessments need to have the following features. They should be:

- Specific: focused on quite narrow content domains so that precise gaps can be identified for future teaching and further practice. Knowledge organisers or personal learning checklists are excellent for helping students to know exactly what they are meant to have learned for any particular micro-test.
- Frequent: building on the idea of regular retrieval practice to develop long-term memory and recall taken from cognitive science (as in Rosenshine's Principles of Instruction featured in Chapter 4).
- Repetitive: to ensure skills and retrieval are actively practised in a focused manner – as happens in sport with skill drills or with rehearsing

⁹⁴ Christodoulou, D. (2016)

a performance where you repeat small elements over and over before assembling the final piece.

- Recorded as raw marks: to ensure the information is kept as close to the details of the original assessment as possible, without being morphed into a bell-curved grade or standardised score.

Summative assessment should be aimed at 'creating a shared meaning'. Here, we need to be sure that our data on assessment has meaning beyond the context of a single classroom so that standards can be compared between classes, subjects and schools. To be effective, summative assessments need the following features:

- Standard conditions: the time, availability of supporting resources, the same kind of questions (obviously enough!).
 - Scaled scores: comparing raw scores across different tests is unhelpful because the level of difficulty can vary between tests; they need to be standardised in a form that allows comparisons to be made.
 - Sampling a large curriculum domain: to ensure that a broad curriculum is maintained and is included in the sampling process that characterises any test design.
 - Infrequent: to ensure that teachers have maximum time for teaching, covering a broad curriculum and engaging in more formative assessment.
- Lots of large-scale summative tests get in the way of the learning process and we don't need that many to provide the tracking information we really need.

Within the area of summative assessment, Christodoulou is keen to stress that different subjects require different specific forms of assessment – as I explore a little further on.

All of this makes perfect sense to me. It has significant implications for the decisions teachers make when planning their assessments. In the Learning Rainforest, the end result will be a sensible blend of formative and summative assessment depending on what is being studied, where students are in their learning and what their particular needs are. Crucially, we need to be clear of the purpose and if we really want students to improve we need to be sure that actions are taken as a consequence of the assessment, whatever form it takes. In practice things like knowledge organisers (see C15 *Specify the knowledge*), planned drills (C18) and regular retrieval practice (K18) are going to be helpful elements in a teacher's Rainforest assessment kitbag.

If these become embedded features of our practice, they allow us to gauge how

well students are doing in terms of how much they actually know about specific topics and how well they can perform specific skills. Assessment is much more powerful when it is formulated purely in terms of what students know and can do in the absence of any proxy grading system or before details are lost in the context of a complex multi-level task or synoptic test. Can you capture all the details or learning in your notebook or data system? No. But why do you need to?

Dylan William: Responsive teaching and feedback

As I described in the introductory chapter, a major influence on me and many others during my career has been Dylan William. When *Inside the Black Box*⁹⁵ by William and Black came out in 1998, the ideas seemed absolutely radical – and yet so logical. One of them was the idea of comments-only marking. It required a major mindset shift for teachers to ditch grades in order to give formative feedback via marking comments that students would actually read – instead of simply reading the grade and nothing more as their research indicated.

Dylan William has continued to be a strong influence and recently, in his conference talks and writing, he has a very clear message. William suggests that, from all the research studies there have been, formative assessment practice encompassing effective feedback is our best bet for raising standards. He says he wishes he had used the term 'responsive teaching' in the very beginning instead of 'assessment for learning' as this conveys the essence of the concept less ambiguously.

For William, it is short-cycle, 'minute-by-minute' feedback that is of particular significance – not the long arc feedback that students might get via marking in books. Short-cycle 'in the moment' feedback – in tight feedback loops – requires teachers to continually engage students in activities that tell them where they are and then, absolutely crucially, to adjust their teaching in response so that students' learning is advanced. This is responsive teaching.

In *Embedding Formative Assessment*⁹⁶ William provides numerous examples of how this works in practice under five main headings – his five strategies. Four of the strategies are:

- Clarifying, understanding, and sharing learning intentions, *eg* devising success criteria for an open-ended task or a successful piece of writing
- Engineering effective classroom discussions, tasks and activities that elicit evidence of learning, *eg* whole-class response methods of questioning

⁹⁵ Black, P. and William, D. (1998)

⁹⁶ William, D. (2011)

- Activating students as learning resources for one another, eg class critique, peer assessment, peer to peer instruction
- Activating students as owners of their own learning, eg self-quizzing, self-assessment against mark schemes or criteria.

The fifth strategy is 'providing feedback that move learners forward'. William explores the nature of feedback extensively in his books and conference sessions. Feedback has been the subject of multiple research studies but it is complex. Feedback can have negative impacts, demotivating students or lulling them into setting lower expectations for their achievement if they reach goals too easily. Also, feedback can be ignored. Given that it can all go wrong, William seems to suggest that good feedback has some particularly important elements:

Primarily, there needs to be trust and good teacher-student relationships. *'In the end, it all comes down to the relationship between the teacher and the student'*. This is so that teachers know their students well enough to judge how best to give them the feedback they will respond to; it is also so that students will listen and respond to the feedback they are given, making the effort needed to effect a change in their performance. In the absence of relationships, these things are much less likely to happen. There's something rather wonderful – and intuitively commonsensical – about this finding: at the bottom of all the technical analysis of feedback, we get down to the need for good relationships. When we're establishing the conditions in our rainforest, relationships based on trust and respect are not just a fluffy feel-good element of a happy classroom: they are a hard-edged necessity for effective formative assessment; for effective responsive teaching; for building knowledge.

In addition to the technical elements of where students should improve, feedback needs to attend to students' motivation. We want students to increase their effort and raise aspirations – and the messages required to achieve those goals can be different from student to student, depending on their past experiences and mindsets. Feedback needs to focus on looking forward, not back; whilst we are learning from mistakes and past performances, the goal is always to improve the next performance. Fairly obviously, there needs to be a next performance or else giving the feedback is a waste of time!

There is then the factor of dependency – SatNav syndrome. If we give too much feedback, students stop thinking for themselves. They need to do the work; they need to learn to self-correct, not simply learn to rely on having their mistakes corrected.

97. William, D. (2016) 'Looking at student work', *ASCD Educational Leadership*, 73(7) pp. 10-15.

William simplifies the complex issue of feedback quality down to this:

The only thing that matters is what the student does with the feedback. If the feedback you're giving your students is producing more of what you want, it's probably good feedback. But if your feedback is getting you less of what you want, it probably needs to change.⁹⁸

Authentic assessment

A key driver of the macro data tracking machinery is the perceived need to communicate the outcomes of assessment in a consistent form and to be able to make comparisons between subjects. There's an assumption that Leo's B in maths indicates better performance than his C in science. However, as reasonable as this might seem, at the granular, cellular level where learning happens, it's not much help at all.

The idea of authentic assessment is that, instead of being driven by the need to produce standardised outcomes, we are primarily focused on generating formative feedback in the form most useful for a particular discipline, whatever that looks like.

In maths and science, there are swathes of assessments where the answers are right or wrong. Most assessments will be straightforward problem-solving or questions requiring simple knowledge recall. This kind of 'difficulty model' assessment, as Christodoulou calls it, is common to lots of subjects. The difficulty of the questions determines the standards and raw scores of tests in discrete topic areas, referenced against a clear knowledge checklist, will tell you what you need to know. It gets a little more complex with multi-mark answers in science or geography, where mark schemes need to spell out a wider range of possible correct answers. The feedback outcomes from difficulty model assessments can be quite specific: the questions that were wrong need to be revisited and attempted again after further teaching and study.

In English and history where extended writing is the most common form of outcome from a learning sequence, comparative judgement might be more useful. These 'quality model' assessments require teachers to compare standards across numerous success criteria (grammar, spelling, depth of analysis, the flow of ideas, the level of originality or perceptiveness). The complexity demands that direct comparison is likely to be required between specific pieces of work and possibly some benchmarking exemplars. Whether or not marks or grades are awarded, these assessments are always based on teacher judgements even if

98. William, D. *ibid.*

tight rubrics and mark schemes are used; they are not purely objective scores as in maths or science.

Feedback is more complex here and teachers need to agree on the language of feedback that helps move students forward. Often this is where things go awry.

- *This paragraph lacks clarity.*
- *The conclusion doesn't address the question sufficiently well.*
- *The ending is too sorry.*

This kind of feedback is too vague to move the learner forward; students would not know what to do differently – only that their work had fallen short. However, with good comments and specific examples – as we examine in K11 Verbal feedback and in the next section – feedback can tell learners much more than any form of grading.

- *Re-write this paragraph using the Point-Evidence-Explain pattern to make it clearer.*
- *In your conclusion, sum up both sides of the argument and then state which you think is the most persuasive.*
- *That 'it was all a dream' ending is over-used. Perhaps try a darker ending that leaves the reader wondering what happened.*

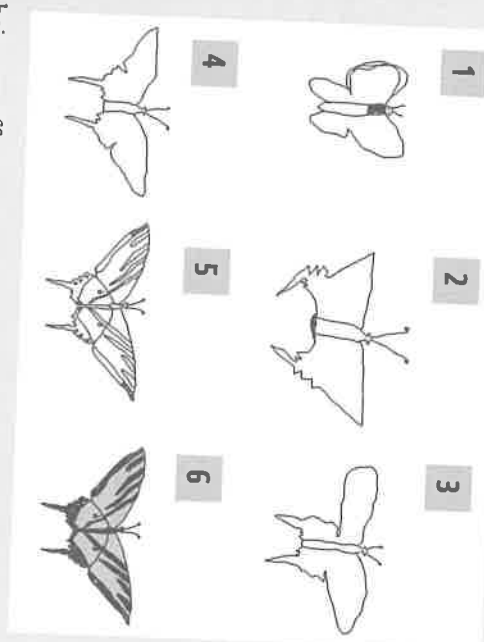
Of course, there is every reason to include simple knowledge recall tests in English and history as part of the diet of formative assessment: dates, characters, the sequence of events, background facts. Knowledge of these elements can be assessed in a raw marks form that is very useful before getting stuck into full-blown essays.

In languages, there are multiple modes of assessment that teachers need to employ to encompass the four skills of speaking, listening, reading comprehension and writing. No single grade can do justice to this range of learning elements and provide information that could be used as feedback. In other subjects such as art, music and drama, there will be some knowledge elements but, largely, the creative outcomes will be subject to teacher evaluation where the interpretation of success criteria will be the key aspect of the process. Are marks appropriate here? Or A-G grades? Or perhaps a simple Pass, Merit, Distinction might be more meaningful given the level of subjectivity. It is self-evident that a B in maths can't mean much in comparison to a B in art in terms of the actual learning so why try to line them up at all? We would do better to keep the information as raw and as close to the requirements for subject-specific forms of feedback as possible.

5.3 Feedback for excellence

BLOG: Lessons from Austin's Butterfly

One of my favourite educational references is the work of Ron Berger and his ideas about cultivating 'An ethic of excellence'⁹⁹. A much-loved metaphor-rich video clip, readily available online¹⁰⁰, is Berger's story of Austin's Butterfly. I refer to this in nearly every training session I do. The video shows Berger telling the story of First Grader Austin and his attempts at 'looking like a scientist' (by which he means using the skill of scientific observation) to draw an accurate picture of a tiger swallowtail butterfly from a photograph. He shows that, by using positive and specific feedback solicited via teacher-led class discussion, through a series of redrafts Austin improves his work from a very basic first effort to a magnificent final drawing.



Beyond being an effective demonstration of the peer critique method, the story conveys several messages. Firstly, there is the idea that Austin was always capable of excellence – it's just that, to begin with, he didn't know what was expected or what he was capable of. But the wonderful final butterfly was in him all along. Linked to this is the idea that, too often, we accept work that is substandard; we associate certain students with

⁹⁹ Berger, R. (2003)

¹⁰⁰ Austin's Butterfly: Building Excellence in Student Work EL Education 2012 <https://vimeo.com/38247060>

their mediocre initial offerings without enabling them to push forward, insisting that they produce and experience excellence. Berger's philosophy is that every child should be given the opportunity to experience excellence:

'After students have had a taste of excellence, they're never quite satisfied with less; they're always hungry.'¹⁰¹

A second layer of thinking is that, in seeking to 'look like a scientist' at a photograph, the end product that would constitute excellence was there for all to see. Everyone involved could use this as a reference point to generate the specific feedback leading Austin towards the goal. From this, we can see that, in our various disciplines, a central task of successful formative assessment is to define our butterflies. Butterflies plural because there may be many ways to achieve excellence.

We need to spell out and exemplify what excellence might look like – including where there are multiple possibilities – and then devise iterative feedback processes that allow students to see the steps from where they are to where they could be in the detail of their learning goals. Austin and his teacher do not need a graded ladder or tracking system to help him improve; it's all about the detail of the work itself linked to a clear idea of what excellence might look like, modelled by an exemplar. Is his butterfly a D or an A? Who cares? It doesn't matter. The value is in the product itself.

Where students are set work that is too easy or where they have substandard work accepted, at the core of the issue is that teachers lose sight of expected standards. They don't define what the finished butterfly equivalent should look like for each learner, taking account of their age and prior attainment. This is often because, in many cases, there is a sliding scale and it is difficult to know.

Take these questions:

- Why does your heart beat faster during exercise?
- What effect does dropping a ball from a greater height have?
- Was it right to kill Osama bin Laden?
- How is Harry Potter different from The Hobbit?
- Was King John a good king?

101. Berger, R. (2003)

It would be possible to have a good discussion about each of these questions with students aged 10, 13, 16 or 18. But what answers should be expected at each level? When should students start talking about haemoglobin and respiration? When should kinetic and potential energy start to feature? What distinguishes an A*/Grade 9 GCSE answer from a strong Year 7 answer about King John, Osama bin Laden or The Hobbit?

Around the world, there will be under-challenged children aged 14 making the same PowerPoint they were asked to make five years earlier; they will still be doing the same kind of percentages and fractions questions that they learned aged 10; they will be having the same discussion about food chains in a forest environment that they had aged 9; they will still be making some teachers happy saying King John was bad and bin Laden deserved to die instead of evaluating competing views in a coherent balanced manner.

At the same time, no teacher deliberately sets under-challenging work. What we need is to do a lot more work on establishing what standards mean, without using a proxy code; we should be doing it by looking at the actual work. Teachers should be discussing this extensively within schools and schools should be sharing this information extensively between them, defining the butterflies.

For any piece of work we should be setting out the most challenging success criteria we can conceive of for the task by referencing specific examples. In Year 8, an exceptional student should be able to produce work like THIS: (produce an actual example). It has the following features: (define the features).

- The minimum expectations we would have might look like THIS: (produce an example; define the features).
- In the 9th Grade, a six-mark answer to a question like 'How does an eye adjust to focus on near and distant objects?' should be structured like THIS with THIS level of scientific detail: (example given).
- Exceptional students in Year 7 might be expected to produce a poem analysis like THIS: (top end example on hand, with features identified).
- By the end of 7th Grade, the highest-performing students should be

looking at solving simple simultaneous equations and Pythagoras' equation in problems like THIS; all students should be comfortable working out any percentage of any number and multiplying and dividing fractions such as THESE...

To me, this is the real scale. Turning all of this real assessment into numbers sucks the meaning out of it. Our obsession with grading is a distraction because knowing that the average for a class is B- doesn't mean much in terms of what they can actually do and what it looks like to achieve excellence. The key to real assessment of this kind, and the real standard-setting that it allows, is to have routine moderation processes. Schools could compile portfolios – writing, questions, work samples, book samples – from top, middle and bottom in each subject and gather in a room with schools nearby, every year.

Subject specialists should pore over each other's evidence of standards and see what everyone else is up to. This would have the effect of making everyone chase the best. Once you see what 13-year-olds can do down the road, you'd be defining your butterfly differently – it would be more and more sophisticated. But every teacher should know very clearly what excellence looks like, for every question they ask and every piece of work they set – not based on what they've always done, but based on sharp current information referenced against tangible exemplars.

How would you measure progress? Through the work. When I worked at the British School in Jakarta, a 3-18 school, every primary child had a 'First of the Month' book. On the first day of every month they would spend an hour writing on a fresh page in that book; the book would travel with them up the school. Over time, you could see their progress simply by turning up the pages. It was wonderful – and powerful as an assessment tool for the teachers. If you turned all that into grades ... I won't go there again.

Let's define standards by looking at the very best examples of work that students can produce – and let's share that information with our students and each other so that our sights are continually being set higher. Let's be very clear about the depth and rigour of the answers we expect students to give at each level in our curriculum so that we're not accepting work they could have produced years ago. It should be a routine part of departmental discourse to clarify expectations of standards, referring to the exemplar material on hand. Let each Austin make the best butterfly

they can ... and not the one they could do already. It should be a matter of basic credibility for any teacher that they stretch the most able in their lessons – there is no excuse not to.

Making marking work as feedback

To conclude this chapter, a brief word about marking. A final part of this macro summative paradigm has been our approach to marking. For years now, marking has been largely a PR exercise where how the marking looks has been more important than what it achieves, with an appallingly low impact to workload ratio. Most marking is wasted. It doesn't lead to students knowing more, understanding things better or producing better quality work. This is in part due to the eternal marking paradox that those students who need the most help are the least able to interpret the real meaning of marking in order to adjust their thinking or performance. It is also partly to do with the time gap between the point when students did the work and the point when they get to improve it: it's often too long and everyone just wants to move to the next thing without making improvements on the last thing.

Despite these issues, school control cultures have demanded levels of marking – of red pen – that bear no relation to students' progress in their learning. As outlined in Part 2 K13 Marking: Keeping it lean, K14 Whole-class feedback and K15 Close the gap, there are alternative methods. Methods that are workload efficient whilst also moving students forward in their learning.

A new assessment paradigm?

If we put all the ideas in this chapter together, we have the ingredients for a more effective assessment paradigm to drive growth in our Learning Rainforest. Is it new? No. But it is certainly better than what dominates now. Instead of macro summative attainment tracking, we want to have a model made up of these elements:

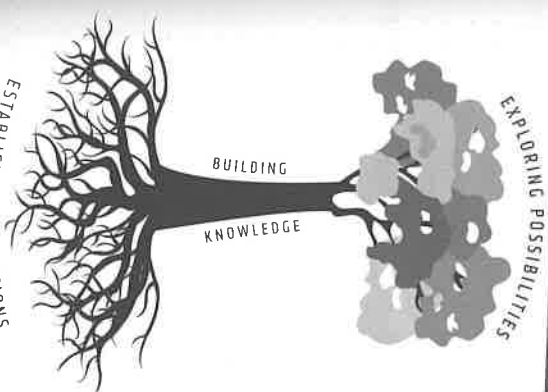
- Maximum use of formative assessment: low-stakes, high-frequency testing owned exclusively by teachers and their students, focused on specific, clearly defined areas of knowledge and skill
- Judicious use of summative assessments that support the leanest possible tracking including the use of comparative judgement wherever subjective quality judgements are being made

- Responsive teaching, where feedback is informed by teacher-student relationships coupled to a range of 'minute-by-minute' interactions that elicit information about where students are in their learning
- Authentic assessment in subject disciplines, where the needs of each subject hold sway over the needs of the centralised data machinery
- Clear exemplars of excellence, where teachers have collectively defined their butterfly, working collaboratively in their subject teams using concrete examples of excellence to develop a shared understanding of standards
- Intelligent, lean marking approaches that acknowledge the limited effect marking has in relation to the demands it places on teachers

MANAGING THE LEARNING RAINFOREST

CHAPTER 6

EACH OF THE SPECIMENS IN THE LEARNING RAINFOREST IS REPRESENTED IN THE WONDERFUL ARRAY OF TREES WITH THEIR COMPLEX ROOTS, TRUNK AND BRANCH SYSTEMS THAT GIVE THEM STRENGTH AND STRUCTURE AND GLORIOUS CANOPIES WHERE THEIR CHARACTER FINDS EXPRESSION.



■ BUILDING THE KNOWLEDGE STRUCTURE:

- USING EFFECTIVE INSTRUCTION METHODS (K1-K10)
- USING EFFECTIVE FORMATIVE ASSESSMENT AND FEEDBACK METHODS (K11-K15)
- TEACHING FOR MEMORY (K16-K20)

■ EXPLORING THE POSSIBILITIES:

- PROVIDING OPPORTUNITIES FOR HANDS-ON, AUTHENTIC LEARNING EXPERIENCES (P1-P10)
- USING ONLINE TOOLS AND OTHER RESOURCES TO SUPPORT STUDENTS AS INDEPENDENT LEARNERS (P11-P16)
- GIVING SPEECH ACTIVITIES A HIGH PROFILE AND CELEBRATING EXCELLENCE IN ALL ITS FORMS. (P17-P20)

■ ESTABLISHING CONDITIONS

- ESTABLISHING THE CONDITIONS, FOSTERING THE ATTITUDES AND HABITS NEEDED AS A BASIS FOR STUDENTS TO ACHIEVE EXCELLENCE. (O1-O5)
- ESTABLISHING EFFECTIVE BEHAVIOUR ROUTINES (O6-O12)
- DESIGNING A CURRICULUM (O13-O20)