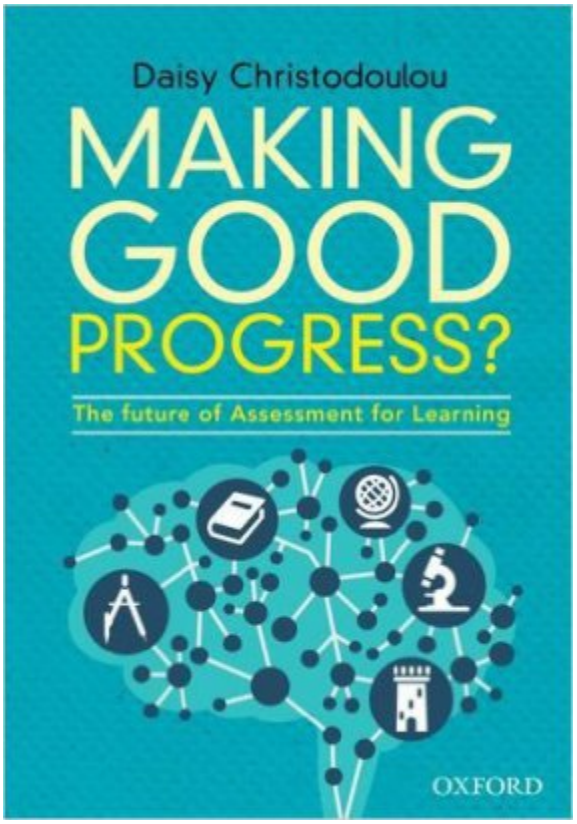
Notes on Making Good Progress – Summary blog



Just because I had written extensive notes, I'd thought I'd just post them in a series of blogs.

Part 1 – foreword, introduction, chapter 1
Part 2 – chapters 2 and 3
Part 3 – chapters 4 and 5
Part 4 – chapters 6 and 7
Part 5 – chapter 8
Part 6 – chapter 9 and conclusion

In conclusion, I think that if a teacher wants to read a timely book with a lot of interesting content on assessment, they do well to read this one. They should, however, read it with the frame of mind that in places the situation is presented somewhat one-sidedly, in my view too negative about the 'old' situation and too positive about alternative models. Teachers can profit from that, but it can also mean that they miss out on decades of unmentioned research on curriculum, psychometrics and assessment. I would therefore encourage them to follow up the references and also read a bit wider. Of course, one cannot write a 1000 page 'accessible' book but given the number of footnotes a bit more depth in some places would have been good. Particular points are:

- Yes, the implementation of Assessment for Learning (AfL) has been problematic. The book covers some on the importance of feedback but not enough prior research is covered.
- I recognise the generic versus specific domain skills discussion but in my view it is presented in a too dichotomous way. There is more than Willingham, for example Sternberg and Roediger on critical thinking. In addition, linking it to leading to certain assessment practices (e.g. teaching to the test) is unevidenced. There also are fair criticisms of deliberate practice.
- The introduction of a quality and difficulty model is useful but again rather binary.
- Reliability and validity are covered but only quite superficially (types of validity, threats to validity etc.), and reliability -in my view- is not covered correctly (the example with 1kg on a scale is an example of reliable AND valid and does not tease out the essential test-retest characteristic of reliability).
- Yes, there are problems with descriptor-based assessments but there is a raft of research addressing their validity and reliability.
- The progression model makes sense but haven't people been doing this for decades? (e.g. in good textbooks).
- Attention given to the testing effect, spaced practice, multiple choice questions is well done.
- Comparative Judgement is worth examining (critically), but (i) no silver bullet, (ii) probably only applicable for niche objectives, (iii) several pressing questions still to ask, (iv) maybe its strength lies even more in the formative realm.

- The proposed integrated system describes what already is in place, with a plea to collaborate. We must realise this is mainly a funding issue, in my view.

One might wonder ‘why mention this, it’s great that it’s mentioned, isn’t it?’ but I simply have to refer to what the author states towards the end of the book. Assessment is a form of measurement and ‘flawed ideas about assessment have encouraged flawed classroom practice’ (p. 212). Great to have made a start in increasing awareness of this, but without covering the basics, those bullet points above, I fear we don’t get the complete picture. I saw some very positive reviews appearing, and I understand why as the book finally gives some well-deserved attention to the area of assessment. However, I also think it’s clear I think the book has some shortcomings; it would be good to keep them in mind while reading it.

Notes on Making Good Progress – the blogs

Sometimes I just get carried away a bit. I managed to get an early copy of Daisy Christodoulou’s new book on assessment called Making Good Progress. I read it, and I made notes. It seems a bit of a shame to do nothing with them, so I decided to publish them as blogs (6 of them as it was about 6000 words). They are only mildly annotated. I think they are fair and balanced, but you will only think so if you aren’t expecting an incredulous ‘oh, it’s the most important book ever’ or ‘it is absolutely useless’. I’ve encountered both in Twitter discussions.

PART 1 | PART 2 | PART 3 | PART 4 | PART 5 | CONCLUSION

PART 1 – THE BEGINNING
This part addresses the foreword, introduction and chapter 1.

I have been following the English education blogosphere for some time now. Daisy Christodoulou might be best known for her book ‘7 myths about education’ (and winning University Challenge with her team). ‘7 myths’ was a decent book with some nice and accessible writing, especially useful because it gave knowledge a bit more attention again. Points for improvement were the fact they weren’t really 7 myths in my view, 3 were variations of another myth, the empirical backing was a bit one-sided, and there was an error in quoting (revised) Bloom. But any way, a fresh voice, and some good ideas; bring it on, now in a new book on assessment.

The **foreword** of the book (again) is by Dylan Wiliam, best known perhaps for his ‘formative assessment’ work with Paul Black. After all the government malarkey on assessment with ‘assessment after levels’ he rightly so emphasises the timeliness of the book. Schools can make new assessment systems. Of course it is telling that a book needs to address this; it could be argued -especially when a government is keen to point at top performing PISA countries- that such an assessment system could be designed by a government. Of course, we now hear this more and more, but only after finishing the old system, opening the way to all kinds of empirically less grounded and tested practices. The foreword ends with a statement I am not convinced by, namely that formative and summative assessment might have have to be kept apart. For instance, it is perfectly acceptable to use worked examples from old summative assessments in a formative way. One could argue that both summative and formative assessments draw from the same source. In fact, in one of the promoted types of

assessment, comparative judgement, one advice seems to be to use exemplars for students to know what teachers are looking for: a summative and formative mix.

One thing that immediately strikes me is that I love the **formatting**. The book has a nice layout and a good structure. Throughout the book, polygon diagrams perhaps suggest more structure than there is (who hasn't used triangles ;-). Contrary to 7 myths each chapter seems to really tackle a separate issue, rather than the same issue in a different guise. The reference lists in the beginning are quite extensive. though for people who know the blogosphere a bit one-sided (Oates, Hirsch etc.). Later chapters have less references, and that is a shame because the second half is far more constructive and less 'this and this is bad' (more on that later). I can agree with a lot of criticisms in the first half, and even with the drawbacks of 'levels', but I am less convinced that some of the proposed alternatives will be an improvement. More evidence would have worked there.

The book starts with an **introduction**. Unfortunately the introduction immediately sets the tone, and in an un-evidenced way. "In the UK, teacher training courses and the Office for Standards in Education, Children's Services and Skills (Ofsted) encouraged independent project-based learning, promoted the teaching of transferable skills, and made bold claims about how the Internet replace memory." I find that a gross generalization. Of course I know about the Robinson's and Mitras of the world, and there probable *are* people in those organisations saying this (and outside), but is it rife? It is a pattern that also was apparent in '7 myths'. The sentence after that with 'pupils learn best with direct instruction' (no, novice pupils, it can even backfire with better pupils, so-called expertise reversal) and 'independent projects overwhelm our limited working memories' (no, this depends on the amount of germane load or, if you will, element interaction) in my view are caricatures of the scientific evidence. Often this has been parried in debates that it is reasonable to simplify it this way. I'm not sure; my feeling is that this is actually how new myths take hold. Luckily, what follows is a good explanation and problem statement for the book; I think it is good to tackle the topic of assessment.

Chapter 1 starts with a focus on Assessment for Learning (AfL). I think the analysis of why AfL failed, partly focussing on the role and types of feedback, is a good one. Black and Wiliam themselves emphasised the pivotal role of feedback, in that it needed to lead to a change in behaviour in the students. This did not seem to happen well enough. On page 21 it is ironic, given what follows in later chapters, that Christodoulou writes "When government get their hands on anything involving the word 'assessment', they want it to be about high stakes monitoring and tracking, not low-stakes diagnostics." I feel that when Nick Gibb embraces 'comparative judgement', this is exactly what is happening. The analysis then continues, on page 23, with sketching two broad approaches in developing skills in the 'generic skills' and 'deliberate practice' methods. I had the well-known 'false dichotomy' feeling here. By adding words like 'generic' and also linking one approach to 'project-based' I felt there clearly was an 'agenda' to let one approach be 'wrong' and one 'correct'. It even goes as far on page 26 to say that the 'generic skills' method leads to more focus on exam tasks. No real support for this supposition. Actually, some deliberate practice methods focus on 'worked examples' where using exam tasks would be reasonable but also 'working with exam tasks'. I agree that approaches should be discussed, by the way, but as so many discussions on the web, not in a dichotomous way if evidence points to more nuance.

PART 2 – SKILLS AND RELIABILITY

This part addresses chapters 2 and 3.

As I think the two approaches in chapter 1 are a bit of a caricature, I wonder whether this continues in Chapter 2. At least there are some good examples of people who, in my view, utilise an exaggerated view of the 'generic skills' approach. Generic skills are rooted in domain knowledge. Yet, it is *not* the case that you will have to re-learn certain skills again and again in every domain. A good example is in my area of expertise maths and spatial/mental rotation skills. There is a (limited) amount of transfer within groups of domains. This is tightly linked to schema building etc. It is therefore unhelpful to present a binary choice here. What *is* good is to make people aware that generic skills courses *need* some domain knowledge. The chapter uses quite a lot of quotes, some from the '7 myths' approach of using Ofsted excerpts. Although I like this empirical element, and I even think there's something in some of the claims, it would have helped if the quotes were something less 'cherrypicking' like. The fact that generic skills are mentioned are no evidence that they necessarily are a focal point of teaching. In fact, generic skills seem to be presented as an almost 'automatic' outcome of 'just teaching' in the deliberate-practice approach. So even in an approach the author seems to prefer, generic skills will probably be mentioned. The chapter of course goes on to name-checking 'cognitive psychology' and Adriaan de Groot (like in Hirsch's latest book). It is good that this research is tabulated, and the addition of 'not easily transferable' already shows a bit more nuance (p. 33). Schemas are mentioned and that it is good that 'acquiring mental models' is put central, and not 'less cognitive load for working memory is best'. I felt these pages showed a wide range of references, though quite dated. I wholeheartedly agreed with the conclusion on page 37 that 'specifics matter' i.e. domain knowledge. It is telling that in discussing this, other 'nuanced' words appear, for example on page 38 when Christodoulou says 'when the content changes significantly, skill does not transfer'. The interesting question then, in my mind, is when content is 'significantly different'. My feeling is that this threshold is often sought far too low by some, and far too high by others. It would be good to discuss the 'grey area', just like the 'grey area' in going from a novice to an expert.

The section concludes with a plea for knowledge, practice etc. with which I very much agree. It becomes the prelude to a section on deliberate practice. It is an interesting section with a role for Ericsson's work. Practice is extremely important; I do wonder, though whether the distinction performance and deliberate practice is more mixed than presented. Originally, the discussion about deliberate practice seemed to revolve around 'effort versus work'. This meta-review suggests there are more things in becoming an expert. Yes you practice specific tasks but I think it is perfectly normal to early on also 'perform', whether it is a test, a chess game or a concert. Or even look at an expert and see how they do (mimic). Not with the idea that you instantly become an expert but the idea that it all contributes to your path towards *more* expertise and solidify schema. Especially the link to 'performance' being too big a burden on working memory does not seem to be supported by a lot of evidence. It is not true that you can't learn from reflecting on 'performance' as many post-match analyses show. Of course, one reason for this might be that the examples all are from 'performing arts' and sports, arguably more restrained by component skills leading to the 'performance skill', but after all it's not me introducing these examples. At the bottom of page 41: "even if pupils manage to struggle through a difficult problem and perform well on it, there is a good chance that they will not have learnt much from the experience." in my view plays semantics with the word 'difficult problem'. I wonder why 'there is a good chance' this is the case. It also

poses interesting questions regarding falsifiability, after all if a student does well on a post-test and has a big gain in an experimental research setting, maybe they haven't learnt anything? Maybe they just performed well. By now, I have seen enough from the over-relied on Kirschner, Sweller and Clark paper. Bjork's 'over-learning' is an interesting addition: I would agree it can be good to over-learn but unfortunately there is no mention of expertise reversal, worse performance. On page 42 and 43 I thought we would get to the crux (and difference) in aims of tasks, because I agree that those are key in learning. While acquiring mental schemas the cognitive load does not have to be minimal, just as long as those schemas are taught. In assessments you don't want the cognitive load to be too high because you will fail your assessment. The chapter finishes with an 'alternative method' as a 'model of progression'. I am not sure why this is called an 'alternative' because it sounds as if it has been around for ages. It even echoes Bruner's scaffolding (oh no!). The attention to peer- and self-assessment is interesting, but I'm not sure if direct instruction methods really incorporate them, at least not in the often narrow terminology used in the edu blogosphere. Although I have seen a broadening of the definition through 'explicit instruction'. I'm sure some will

point out, that oft ridiculed progressive behaviour, of not understanding the definitions
In sum, a useful chapter with a bit too much of a false choice.

The start of chapter 3 puzzles me a bit because it starts by explaining how summative and formative functions are on a continuum. I agree with that, and find it at odds with Wiliam's foreword, in which he seemed to confess that the functions need to be separated. The chapter discusses the concepts of validity and reliability. I am not completely sure I agree with the formulation that validity only pertains to the inferences we make based on the test results, but I haven't read Koretz. There are many types of validity and threats to validity, and I would say it *also* is important that a test simply measures what it purports to measure (construct validity); the many sides of the term should be discussed more. The comment on sampling is an important one. With reliability, I think the example with a bag of flour of 1 kg is an awkward choice, as it suggests a measure can only be reliable -in this case- as it shows 1 kg. This is not the case, scales that consistently measure say 100 grams over would still be a reliable scale, just not valid for the construct measured (mass). Reliability also isn't an 'aspect of validity'. When discussing unreliability it would have been helpful to have been more precise with explaining the 'confidence bands', and perhaps measurement errors. I get the feeling that the author wants to convey the message that measurements often are unreliable, but maybe I'm wrong. I very much like the pages (p. 64) on the quality and difficulty model; I agree that both models are accompanied by a trade-off between validity and reliability. There is a raft of literature on reliability and validity, Christodoulou chose only a few. As a whole, the chapter makes some useful links with summative and formative assessment. However, the example on page 70 is not chosen very well (and again note that there are many long quotes from other sources, more paraphrasing would be helpful), as in my view the first example (5a + 2b) *can* be a summative question if pupils are more expert (e.g. maths undergraduates). I like how Christodoulou tries to combine summative and formative assessments in the end, but wonder what new baggage we have learnt to make that happen.

PART 3 – DESCRIPTOR- AND EXAM-BASED ASSESSMENT
This part addresses chapters 4 and 5.
Chapter 4 critiques descriptor-based assessments. I think it is important here to distinguish a bad implementation of a good policy or simply a bad policy. It starts by describing 'assessment with levels'. I notice that the author often takes reading examples, which in

principle is fine, but the danger is that we too quickly think it applies to all subjects. I think the chapter does a good job at describing the drawbacks of descriptor-based systems. I do, however, feel that some of them are not less prominent in alternatives presented later. I also get the feeling that apples and oranges are sometimes compared in the 'descriptive, not analytic' section because there is no reason to not simply do both. The comment on 'generic, not specific' with regard to feedback is spot on, but again there is no reason to not then do both: generic AND more specific feedback, in my opinion. Actually, throughout the book I feel that the novice/expert cut that had so skillfully been exposed, is not taken into account in many of the pages. As reviews of feedback use have shown, the type of feedback (and timing) interacts with levels of expertise. The examples of different questions seem related to their specific goal e.g. on page 94 the question on Stalin can be an excellent multiple choice question on certain knowledge. However, if it was more about relationships of certain events multiple choice questions might give away the game too much. The same with equations: multiple choice questions do not make sense if your aim is to check equation solving skill, but would make sense if you want to check if they can check the correctness of solutions. I think there is some confusion about reliability and validity here, most prevalent in the example on fractions. Yes, the descriptor on fractions is general but that is often part of a necessarily somewhat vague set of descriptors in a curriculum. What Christodoulou then gives as example (page 99) seems to be more about validity and reliability of tests and assessments. Decades of psychometric research have provided insight in how to reliably improve assessment for summative purposes. It feels as if this is under-emphasised. Also, descriptor systems can be made more precise by mark-schemes and exemplars (as, by the way, later on presented in the comparative judgement context). A pattern in the book seems to be that

1. The author provides some *good critiques of the drawbacks of existing practices*,
2. But then does not mention *research* on mitigating drawbacks,
3. Nevertheless, *a case is made for changes* with a 'solution'
4. But these solutions are not discussed in light of how they improve the drawbacks and/or introduce other drawbacks.

This could lead to a situation where readers might nod along with the critique but then incorrectly assume the proposed solutions will solve them. I think it is admirable to describe the challenges in this accessible way but would have preferred a more balanced approach. As a case in point, take the 'bias and stereotyping' of page 104. This is a real challenge, and rightly so seen as a point to address in descriptor-based assessment. Yet, as said before, there are ways to mitigate these drawbacks. Instead, the case is made that reform is necessary, and later on in the book a 'solution' is given that still uses teacher judgements but 'simpler' (not really, holistic judgement is not simple per se, only if you have a short uni-dimensional judgement to make, but the condemnation of teacher judgement wasn't about that, it was about complex judgements). In my view it just 'pretends' to be a solution for these well-observed challenges.

Chapter 5 critically assesses another assessment type, namely exam-based assessment. The somewhat exaggerated style is exemplified by the first sentence "we saw that descriptor-based assessment struggles to produce valid formative and summative information". The chapter first links the exam model to chapter 3's distinction of the quality and the difficulty model. I am not convinced by the arguments that then try to explain why exam-based (summative) assessments are difficult to use for formative purposes. Sure, they are samples from a domain, but one can simply collect all summative questions on a certain topic or

subject, to make valid inferences. Sure, questions differ in difficulty, but there are ways to analyse the difficulty. The comments on page 120 and 121 are fair (hard to say why right or wrong) but I can't help but think about the 'solutions' provided later on with comparative judgement, which uses Rasch analysis and 'just correct or incorrect', suffer a same problem (granted, they are presented as 'summative' solution). With maths exams there are mark-schemes, so a more fine-grained analysis *is* possible for formative purposes. The chapter *does* provide a nice insight in the difficulties regarding marking and judgement. A third problem, it is suggested, is that marks aren't designed to measure formative progress. I think again that the book asks some good critical questions, but ultimately too much sends out the message that old practices are bad. From page 130 the author argues there are issues with the summative affordances of exams as well. I think that this section, again with the fractions examples, exaggerates the 'non validity' of exams. Testing agencies have developed a raft of tools to make exams valid over years and between samples. Again, the challenges and difficulties are described well, but ways to mitigate the challenges are undermentioned. Further, the suggested 'modular' approach is good but is this really new? The next four chapters are about alternative systems.

PART 4 – A MODEL OF PROGRESS AND PRINCIPLES OF ASSESSMENT
This part addresses chapters 6 and 7.

Chapter 6 describes the first of the alternative models, with the model of progress. I think it makes perfect sense to link summative and formative assessments, and I also applaud the suggestion that textbooks, or even digital textbooks, could play a larger role in the English curriculum. Here, I have been influenced by my Dutch background, where using textbooks (for maths, for example, my subject) is quite normal. There also is ample research on textbooks from other countries. 'Progression' also seems to refer to starting with basic skills and 'progressing' to next phases. I'm immediately thinking about (sequences of) task design, worked examples, fading of feedback, scaffolding, etc. These are all common elements of instructional design and multimedia learning and remain unmentioned. I think it's good that the idea of 'progression' is made accessible for the average teacher but do wonder whether this is a missed opportunity. In designing their lessons teachers can be helped, even for the domain of assessment. It is followed up by some interesting threats to validity, including teaching to the test. I thought the author's description of a progression model makes sense; I imagine it is what humans have done over the centuries while designing curricula. Measuring the progression (p. 155) repeats the assumption that if you are interested in generic skills (I agree that with Christodoulou that's not enough) you will grade frequently. In my mind it seems a bit of a rhetorical trick to make generic skill lovers complicit to a testing regime. It is interesting that Christodoulou mentions the word 'multidimensional' because I will later on see it as one of the summative shortcomings of comparative judgement, which promotes an holistic judgement over separate elements Of course I agree with the advice we "need different scales and different types of assessment" (p. 159) and I also like the marathon analogy. But I do wonder what is new about that advice.
Then it's onwards to principles for choosing the right assessments in chapter 7. To improve formative assessments some elements are promoted: specificity, frequency, repetition, and recording raw marks. I like how multiple-choice questions are 'reinstated' as being useful. I do think the advantages are exaggerated, especially because the subject context is disregarded, as well as multiple-choice 'guessing' strategies. It is notable that Christodoulou goes into the latter criticism and explains how the drawbacks could be mitigated. I think it would have been good if these had also been addressed for more subjective essays. The maths example of p. 167 is fair enough, but there technically (even with marking) is no reason to not make this

an open question that can even provide feedback. I think it also would be useful to distinguish between different types of knowledge that should underpin questions. I think it is perfectly fine to give MC questions a firm place for diagnostics (or even diagnostic databases, as there already are many of them) but the author could highlight cutting edge potential more as well. Maybe it's most useful to simply not say that one or the of the question type is 'best suited' but to simply say one needs to ensure that the inferences drawn from the questions are valid; in other words the validity of them. 'Validity' seems to be a term that underpins a lot of the author's thinking, which makes it a shame that it wasn't treated more elaborately in chapter 3. I like how the testing effect, and Roediger and Karpicke's work, features from page 169, as well as desirable difficulties (Bjork) and spaced and distributed practice. These are all very relevant and indeed could inform teachers how to better organise their assessments.

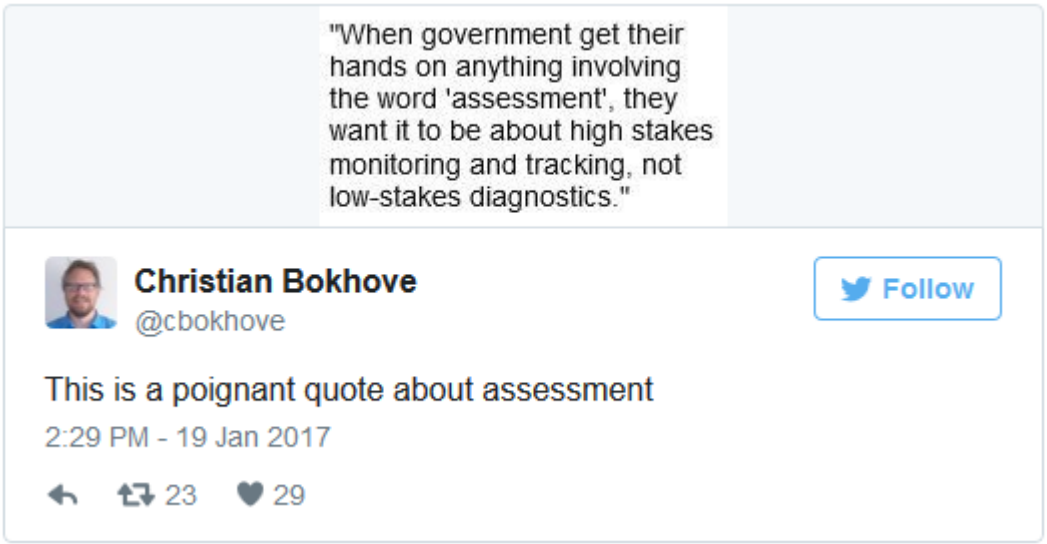
PART 5 – COMPARATIVE JUDGEMENT AND CURRICULUM-LINKED ASSESSMENTS
This part addresses chapter 8.

Chapter 8 addresses the topic I started out reading this book in the first place: improving summative assessments through comparative judgement (CJ). This previous post, which I wrote right after reading this chapter, asks some questions about CJ. The chapter starts by repeating some features for summative assessments. The first is 'standard tasks in standard conditions'. But it isn't really about that in the subsequent section, but 'marker reliability' (p. 182). The distinction between the previously described difficulty model and quality model, is useful. It is clear to me that essays and such are harder to mark objectively, even with a (detailed) rubric. It is pertinent to describe the difference between absolute and relative judgements. However, when the author concludes "research into marking accuracy…distortions and biases" she again disregards ways to mitigate these issues, even while the referenced Ofqual report does mention them. Indeed, many of the distortions are 'frustrating' judgements, and therefore a big danger of rubrics. I, however, find it strange that this risky point of rubrics is disregarded, when comparative judgement, it is suggested, can work with 'exemplars'. As Christodoulou pointed out on p. 149 there is a danger that students work towards those exemplars. I saw this often in some of the Master modules I taught: a well-scoring exemplar's structure of sub-headings was (awfully) applied by many of the students, as if they thought that adopting those headers surely had to result in top marks. So in a sense I agree with the author's critique, I just don't see how the proposed alternative isn't just as flawed. Then comparative judgement is described as very promising. It is notable that most examples of its affordances feature English essays. It also is notable that 'extended writing' is mentioned, while some examples are notably shorter. The process of CJ is described neatly. I think the 'which one is better' is glanced over i.e. 'in what way'? I also think more effort could have been put in describing the algorithm that 'combines all those judgements, work out the rank order of all the scripts and associate a mark for each one' (p. 187). The algorithm is part of the reason why reliability is high: inter-rater reliability can be compared with regard to rank orders; I am not sure if the criticised traditional method is based on rank orders. For instance, if one rater says 45% and another 50% it seems reasonable to say that both raters did not agree. Yet, if we just look at the rank order they might have agreed that one was better than the other. As CJ simply looks at those comparisons, reliability is high. But it's not comparing like with like. A similar process with one marker and 30 scripts would involve ordering scripts, not marking them.I have to think about several challenges that are mentioned in this older AQA report. I don't think these challenges have yet been addressed nor discussed.

I think it is also interesting that Christodoulou correctly contrasts with (p. 187) ‘traditional moderation methods’. Ah, so not the assessment method per se, but the moderation. The Jones et al. article is referenced, but the book fails to mention how the literature also mentions several caveats e.g. multidimensionality and length of the assessment. The mentioning of ‘tacit knowledge’ is fine but it is not necessarily tacit knowledge that improves reliability, in my view. It can be collective bias. I think it’s a far stretch to actually see the lack of feedback for the scripts as an advantage, because it ‘separates out the grading process from the formative process’. It even distributes the grading process over a large group of people; to a student it can be seen as an anonymous procedure in the background. Who does the student turn to if he/she wants to know why he/she got the mark she received? Sure, post-hoc you can analyse misfit, but can you really say -as classroom teacher- you ‘own’ the judgement? Maybe that is the reason why it is seen as advantage, but one can rightly so say the exact opposite. It is interesting to note that the Belgium D-PAC process actually seems to embrace the formative feedback element CJ affords. The section ends with ‘significant gain of being able to grade essays more reliably and quickly than previous’. I think the ‘reliably’ should be seen in the context of ‘rank ordering’, length of the work, and multidimensionality. ‘Quickly’ could be seen in more than just the pairwise comparisons (it is clear that they are short, if ‘holistic’ is only needed); but the collective time needed often surpasses the ‘traditional’ approach. ‘Opportunity cost’ comes to mind if we are talking about summative purposes through CJ. I am disappointed that these elements are not covered a bit more. The section, however, ends with what I *would* see as one big affordance of CJ: CJ as way to CPD and awareness of summative and formnative marking practices. But this is something different than a complete overhaul of (summative) assessment, because of the limitations:

- Needs to be a subjective task (quality model, because otherwise there are more reliable methods)
- Can’t be too long (holistic judgement would most probably not suffice)
- Can’t be multidimensional (holistic judgement would most probably not suffice)

That’s quite a narrow field of application. And with the desire to stay in the summative realm, in England, only summative KS2 not-too-extended writing seems to be the only candidate (see for formative suggestions the previous blog on CJ). But careful:



https://twitter.com/cbokhove/status/822088633985486848

In my opinion, page 188 also repeats a false choice regarding rubrics, as the described ‘exemplars’ can also be used with rubrics, not only with CJ. We do that in aforementioned

Masters module (with the disadvantage it becomes a target for students). So although I agree this would be 'extremely useful', it actually is not CJ. Another unmentioned element is that CJ could be linked to peer assessment. To return to page 105 where bias is seen as human nature, one could argue that a statistical model is used to pave over human bias. In my opinion, this does not mean it's not there, it's just masked.
The second half of the chapter addresses curriculum-linked assessments. I don't understand the purpose of mentioning GL assessment, CEM and NFER apart from using the unrealistic nature of using them to argue 'we need something in-between' summative and formative, and then to argue 'curriculum linked'. As previous chapters, good points are raised but it feels the purported solutions aren't really solutions; the problems are used to argue *something* must change but not so much why the suggested changes would really make a difference. For example, the plea for 'scaled scores' is nice but I would suggest only people who know how to deal with them, should use scaling; simply applying a scaling algorithm might also distort (think of some of the age-related assessments used in EEF studies, or PISA rankings).

PART 6 – INTEGRATED MODEL AND CONCLUSION
This part addresses chapter 9 and the conclusion.
Finally, chapter 9 tries to tie several things together in one 'integrated assessment system'. There are no references in this chapter. Many elements have already been discussed. For example, the 'progression model' which did not seem to offer really new insights (at least to me). Lesson plans and schemes of work appear out of the blue, together with 'curriculum'. I agree that textbooks would be most helpful here. Another element is a 'formative item bank'. Again, very useful, and there are already plenty out there. I am not sure if the summative item bank would need to be a different bank, just the way the items are used and compiled in valid, rigorous summative assessments needs scrutiny. I felt the 'summative item bank' for the quality model was far too much geared towards comparative judgement, an approach that in my view has limited scope; descriptor-based assessments can still play a role, especially in relation to exemplars. What the model *does* emphasise is that an assessment system should draw from several summative and formative sources, perhaps a little bit contradicting earlier parts of the book. This is also expressed on page 206 with the benefits (coherence, pupil ownership with adaptivity and gamification, self-improving with more adaptivity). Ultimately, though, I am left with the feeling that all these elements are already readily understood and even in place. Christodoulou seems to realise and state this on page 207 "Every individual element of this system exists already", but does not address *how* organisations could come to an 'unprecedented collaboration'. Maybe the challenge is that so many people *have* already tried and failed. Ideally I would have wanted the author to have touched on the costs for the resources as well. Many item banks cost money, GL and CEM assessments cost money, No More Marking is not free, Textbooks and exam boards charge money. All in all, with a funding squeeze, it is unrealistic to not address the costs.

The conclusion in the book is rather meagre with three pages. There is some repetition and bold claims again 'flawed ideas about assessment have encouraged flawed classroom practice'. I think this caricaturises the situation. Sure, there are flawed practices but one could also say -in the quest for valid and reliable assessments- there always are flaws, even in some of the solutions Christodoulou proposes. Rather than exaggerate by calling practices flawed, it is better to look how practices can be improved. Christodoulou has some suggestions that should be taken seriously, but also critically evaluated in light of the wide body of research on assessment.