Witness Name: Dennis Sherwood
Statement No.: **M08-SHERWOOD-001**
Exhibits: **DHS/01 - DHS/03**
Dated: 18 August 2025

UK COVID-19 INQUIRY

WITNESS STATEMENT OF DENNIS SHERWOOD

Preamble

I, Dennis Sherwood, will say as follows: -

1. I am, by profession, a management consultant. I was a consulting partner in Deloitte, Haskins + Sells, and also Coopers & Lybrand, for 12 years, and since 2000, I have been running my own consulting business. In 2013, I was commissioned by Ofqual, the regulator of school exams in England, to study their systems. That sparked an interest in examinations and assessment, an interest I have maintained ever since.

2. Accordingly, during 2020 and 2021, I was a close observer of how, in England, GCSE, AS and A-level grades were being determined in the absence of formal examinations. During that time, I authored 23 articles and blogs, published on platforms such as the Higher Education Policy Institute (HEPI), Times Educational Supplement (TES), Schools Week and LSE. The first is dated 21 March 2020, just three days after Boris Johnson's announcement that exams were to be cancelled; the last is dated 7 August 2021, a few days before that year's A-level results (Sherwood, 2020 and 2021). In addition, I made seven submissions to the House of Commons Education Select Committee, as well as being in regular contact with many journalists, providing them with information and leads. Later, during 2022, I wrote a more complete account of what had happened, as recorded in Chapters 10 to 13 of my book, *Missing the Mark – Why so many school exam grades are wrong*, published by Canbury Press in August 2022 (Sherwood, 2022).

3. I was also a member of the Facebook group known as 'ALGI', a support community for A-level students and parents directly affected by the events of summer 2020, and so the 'recipients' of the outcomes of that year's process; ALGI then continued into the following year. Since my own two children had long since grown up, unlike every other member of ALGI, I was neither an A-level student nor a parent of one. It so happened that, in 2018, I had met the Group's key organiser, Elaine Hughes, who knew of my interest in examinations. I was therefore invited to join the Group in a 'peripheral' capacity, providing information from time to time.

4. I note that my knowledge, observations and opinions – and hence the content of this witness statement – relate to what happened only as regards GCSE, AS and A-level examinations in England. However, given that the processes used for the equivalent examinations across the four nations were similar, many of the observations I make are quite likely to apply to Scotland, Wales and Northern Ireland too. And let me stress that those observations, and the corresponding opinions, are mine, and mine alone.

5. I trust that this document is clear and complete in its own right without recourse to further material; that said, I attach two exhibits, and give references to all my sources.

The impact of the cancellation of exams on children

6. I believe that many students, when informed of their confirmed summer 2020 and 2021 GCSE, AS and A-level grades, would have felt both pleased and relieved: that's because, as shown in the following table, the percentages of students awarded high grades in England in summer 2020 were substantially greater than in the preceding years, and even more so in summer 2021, especially for A-level:

	A-level		AS		GCSE	
	Grades A* and A, % of total	Total number of grades awarded	Grades A and B, % of total	Total number of grades awarded	Grades A* to C, and 9 to 4, % of total	Total number of grades awarded
2017	26.2%	759,233	43.1%	635,175	66.1%	4,970,087
2018	26.2%	745,537	47.2%	260,710	66.6%	5,013,364
2019	25.2%	736,734	37.4%	114,088	67.1%	5,075,675
2020	38.1%	718,857	49.4%	70,550	75.9%	5,214,037
2021	44.3%	752,554	50.2%	56,559	76.9%	5,236,861

Source: Joint Council for Qualifications, JCQ

Table 1: Percentages of higher grades, and total grades awarded, 2017 to 2021

7. Since 2012, a primary policy of the exam regulator Ofqual had been to control 'grade inflation', this being the term used to describe the gradual year-on-year upwards trend in the percentage of top grades, as had been happening throughout the 1990s and 2000s. Ofqual's success in implementing this policy – with the exception of 2018 AS, which, historically, is an outlier – is evidenced by the figures for 2017, 2018 and 2019 as shown in Table 1. Had 'real' exams been held in 2020 and 2021, Ofqual would surely have maintained this policy, and the percentages of top grades would certainly have been around 26% for grades A* and A at A-level, 40% for AS grades A and B, and 67% for grades 9 to 4 at GCSE. The higher figures for the 2020 'Centre Assessed Grades' ('CAGs') and 2021 'Teacher Assessed Grades' ('TAGs') therefore imply that many students received higher grades than they would have done had they sat 'real' exams. This opened doors for progression, especially since universities and colleges honoured their offers and accepted larger cohorts than they had originally planned for.

8. Some students awarded higher grades might have worried that the greater number of top grades might have 'devalued the currency', fearing that, in the years to come, a prospective employer might say "You got your A-level grade A in 2020, when everyone got As. So your grade A is only as good a 'real' grade C". That is a possibility, but in my opinion, a remote one. Yes, in 2020 and 2021, more young people were awarded higher grades than in the late 2010sm – and also, as we now know, than in the subsequent years 2022, 2023 and 2024, over which Ofqual have ensured a return towards the historic grade distributions of the years before Covid-19. The 'good news', though, is that these higher grades offered opportunities – opportunities from which, I trust, those

students largely benefited. The Inquiry, however, may wish to investigate whether or not current users of grades, especially employers, are accepting 2020 and 2021 grades 'at face value'.

9. Not all students, of course, were awarded higher grades, and some who did not receive top grades might still have been disappointed. Some might have accepted their awarded grades with good grace; others with shrugged shoulders. But there were undoubtedly some who became frustrated or angry. Frustrated because, according to Ofqual's rules, in 2020, the CAGs could not be appealed at all, whilst in 2021, the TAGs could be appealed only on the grounds that the academic judgement used to determine any TAG had been 'unreasonable', which was very hard for an aggrieved student or parent to demonstrate.

10. And angry not only at the 'brick walls' erected by schools to justify their judgements, but also at the opacity of the processes by which the CAGs and TAGs had been determined; anger aggravated by the possibility that the same student might have been awarded a higher CAG or TAG had he or she attended a different school, and so been 'judged' by different people – people who might have had a different (and perhaps more favourable) view, people who might not have held some form of bias, from "I never liked that kid" to racial prejudice.

11. 'Real' exams have the benefits of clarity and objectivity of process, in that all students take the same exam at the same time, the scripts then get marked, and finally grades are assigned. There is also 'double anonymity': the student doesn't know the marker, and the marker doesn't know the student, so avoiding the possibility that 'teacher's pet' might get a high grade.

12. The 2020 CAGs, and 2021 TAGs, however, failed to comply with these principles. The processes by which both the CAGs and the TAGs were determined were not only unclear, but, in some unknown way, were determined by the inevitably subjective judgement of those who knew the students well, their teachers. Furthermore, different schools executed whatever-the-process-was differently – for example, in 2020, one school might have taken the results of 'mock' exams into consideration, but a neighbouring school might not have done, perhaps because a mock had been planned for a date after the lockdown, and so did not take place. To make matters worse, the 'audit trail' of evidence as to why a student had been awarded [this] CAG rather than [that] one could never be as robust as the student's answers to questions 4, 5 and 6 in

a national sit-down, paper, exam. And with the judgements being made by the students' teachers, 'Alex', an aggrieved student, might well believe that 'Sam' was awarded a higher CAG or TAG simply because the teacher liked 'Sam'.

13. Another, and most important, difference between the 'normal' process, and the processes of 2020 and 2021, concerned appeals. After a 'real' exam, any student unhappy with a grade can raise a 'challenge', causing the exam board – the 'arms-length body' who sets and marks the exam – to carry out a 'review of marking', with the school acting only as the conduit between the student and the exam board. In 2020 and 2021, however, the exam boards stood aside, for they had played no role in the determination of the CAGs and TAGs. Any challenge of a CAG or TAG was a matter between the student (or rather, in many cases, the student's parents or guardians) and the school. This led to any number of bitter, long-running, draining disputes as schools defended their positions, and parents became increasingly irate, seeking reasons why what had happened was not merely an academic misjudgement but malpractice, bias or discrimination, these being legitimate and recognised grounds for appeal. As many of the ALGI members will verify, these disputes – some of which remain unresolved to this day – soured relationships between the school and the parent, with potentially damaging consequences to any younger siblings at the same school.

14. Anger was especially virulent as regards the process associated with the 2020 CAGs. Whereas the 2021 TAGs were determined solely by local teacher judgement, the 2020 CAGs were determined by teacher judgement in the context of what teachers could infer about the mysterious process known as 'statistical standardisation', as implemented by what the then Prime Minister, Boris Johnson, subsequently referred to as the "mutant algorithm" (Prime Minister's Office, 2020). Ultimately, the algorithm itself had no impact on students directly, for its 'calculated grades' were scrapped. There is, however, no doubt that it had a most significant indirect impact, influencing how teachers allocated CAGs to their students, and, as is quite likely, distorting their judgement.

15. When the AS and A-level results were announced on 13 August 2020, the furore that erupted across the media inflamed public opinion, not only discrediting Ofqual, but also contributing to one of the most negative aspects of that year's outcomes – destroying trust in teachers. This had a highly damaging consequence – the very poor process adopted in 2021, in which Ofqual and the exams board chose to stand aside, leaving everything to the teachers. Who therefore took all the blame.

16. This is of relevance right now, for as I write this – in June 2025 – the Curriculum and Assessment Review, currently being undertaken under the leadership of UCL Professor Becky Francis, is considering options for the future of education, options that could well have a decisive influence on the policies to be adopted over the coming decade or more. Within the Review's terms of reference is the balance between formal exams and teacher assessment. However, given the damage done to the credibility of teacher assessment in 2020 and 2021, it is no wonder that the Review's interim report drops a strong hint that exams, as we know them, are here to stay: "*Externally set and marked exams are an important way to ensure fairness as part of our national qualification system*" (Curriculum and Assessment Review, 2025).

17. In my opinion, those who took the decision in 2020 to use the algorithm, and the associated process, made a grievous and immensely damaging error – especially since much more practical, and far more robust, alternatives were possible (Sherwood, 2020a). And even more so since, at that time, Ofqual had full knowledge that, far from "*exams being fairest way to assess student performance*" (Williamson, 2020), 'real' exams grades were in fact only 75% reliable, in that about 1 grade in every 4, as awarded in the 2010s (and as still being awarded now), was (and continues to be) wrong (Sherwood, 2019a), as will be discussed in more detail in paragraphs 183 to 209. Surely well-supervised teacher assessment would have given results more reliable than the 'real' exam 'benchmark' of 75%, and in paragraphs 217 to 234 I present some evidence suggesting that the 2021 TAGs did indeed do this, perhaps achieving about 90% reliability.

18. Despite the magnitude of the 2020 disaster, only two of those involved – Sally Collier, Ofqual's Chief Regulator, and Jonathan Slater, the Permanent Secretary at the Department for Education (DfE) – 'departed'. The two most senior people, the then Secretary of State, Gavin Williamson, and the Minister of State for Schools' Standards, Nick Gibb, were ultimately rewarded with knighthoods in 2022 and 2025 respectively; one of the Ofqual Board Members, Ian Bauckham, who was present throughout this time and, presumably, in agreement with what was happening, was also knighted (in 2023), as well as being appointed Ofqual's Chair in 2021, and subsequently, as of January 2024, Chief Regulator.

19. The Inquiry may wish to examine the circumstances in which this fateful decision was taken, and why, throughout the summer of 2020, Ofqual refused to listen to those many people – including the House of Commons Education Select Committee – who sought

clarity as to how, precisely, the algorithm would work, and who expressed concern about the possibility that the results might not be sufficiently accurate or statistically sound, as well as potentially biased against, for example, ethnic minorities and students with special educational needs. Should the Inquiry wish to do this, I trust that paragraphs 20 to 159 will be of interest. Paragraphs 160 to 178 present a brief description of the (much simpler) process associated with the summer 2021 TAGs. I end this submission with some more general observations (paragraphs 179 to 216), and an exploration of the possibility that, despite all that has been said, the CAGs, and more so the TAGs, were perhaps more reliable assessments than written exams (paragraphs 217 to 234).

2020 – The context

20. To set the events of 2020 in context, I highlight here two important features of the school exam system, as it evolved during the 2010s, primarily under the political influence of Michael, now Lord, Gove, and his then Special Advisor, Dominic Cummings.

21. The first is the fact that, by the late 2010s, students' grades were determined almost exclusively from final exams, and only final exams. If something should happen to those exams – such as a major security leak, let alone a pandemic preventing the assembly of students in the exam hall – then the whole edifice crashes, for there is no other source of reliable information on which to draw, such as appropriately marked coursework or intermediate exam results. The summer exams are therefore what engineers refer to as 'a single point of failure', and, in 2020, fail they did. It is possible that the authorities – Ofqual, the DfE, the exam boards – had a well-thought-through contingency plan, but if they had compiled one, it had not been made public and agreed by teachers. Certainly, in March and April 2020, as the details of what needed to be done emerged, it seemed as if the process of CAGs, rank orders and the use of a mysterious algorithm had been invented off-the-cuff.

22. The second systemic feature is far deeper and much more problematic. Faced with the cancellation of exams, it does not take too much imagination to come up with the idea that teachers might be in the best position to determine students' grades, for teachers, in general, know their students well. But for that to happen, teachers have to be trusted. Trusted by their students. Trusted by parents. Trusted by the authorities. And that's the fundamental problem. Although the authorities will never admit it, my belief is that the authorities did not trust teachers at all, for that mistrust underpins many of the official

acts taken during the 2010s, such as the significant reduction of teacher-judged coursework which was allowed to contribute to students' assessments, as well as official statements such as "*exams are the fairest way to assess student performance*" (Williamson 2020). Had there been a glimmer of trust in teachers, the process in 2020 would have had no need for an algorithm, and all the effort would have been focused on ensuring fairness and a consistency of standards across the country, as, in my view, would have been quite feasible by well-organised peer review. Yes, after the chaos of 2020, teacher judgement was used for the 2021 TAGs, but the process Ofqual used then was so weak that you don't have to be a cynic to surmise that teachers might have been deliberately set up to fail.

23. But suppose for a moment that teacher judgement were to have been trusted in 2020 when there were no formal exams. Had that happened, that raises the question "if we can trust teacher judgement when there are no exams, why do we need exams at all?", or, in a 'softer' form, "why do we place so much reliance on final exams when we can have a 'hybrid' system in which a candidate's grade is determined [this much] by teacher judgement and [that much] by formal exam?".

24. Such questions, of course, are anathema to those with vested interests, such as the exam boards, who make money selling exams, and Ofqual, whose existence depends on them. And there's political dogma too. In an era of policy driven by the likes of (now) Lord Gove and (now) Sir Nick Gibb, the possibility that teachers might be trusted was heresy indeed.

25. These systemic failures yielded their fatal fruit in 2020 and 2021. The agreed, and well-established, contingency plan should have been to trust teacher judgement; trust built up over the preceding years by virtue of the contribution of teacher judgement to students' final grades in the 'normal' process. If we wish to prevent the chaos of 2020 and 2021 happening again, that's how to do it.

26. Within this overall context, there were a host of problems associated with the 2020 process, as described in the following paragraphs:

- 'Statistical standardisation' and the algorithm (paragraphs 27 to 29).
- CAGs, rank orders and the Head's declaration (paragraphs 30 to 34).
- Why the CAGs and rank orders were not needed (paragraphs 35 to 43).
- Misleading information (paragraphs 44 to 47).
- The 'history trap' (paragraphs 48 to 53).

- The fact that different schools used different processes (paragraphs 54 to 58).
- The statistical interpretation of historic data (paragraphs 59 to 62).
- The need to round fractions and decimals to whole numbers (paragraphs 63 to 82).

2020 – 'Statistical standardisation' and the algorithm

27. Fundamental to the award of the summer 2020 GCSE, AS and A-level grades in England was a process referred to as '*statistical standardisation'*, described in Ofqual's 'Guidance' of 3 April as the use of *"a statistical model to standardise grades across centres in each subject"* (Ofqual 2020a, page 11). The reality, and truth, is that this was the application of a computational algorithm to determine, for each school, how many students were to be awarded each grade in each subject. So, for example, for the 28 students taking GCSE Geography at 'Ambridge High School', the algorithm allocated two grade 9s, two grade 8s, three grade 7s, eight grade 6s, six grade 5s, three grade 4s, and one for each of grades 3, 2, 1 and U. These grades were then associated with specific individual candidates in accordance with the student rank order submitted by the school, as discussed further in paragraph 33.

28. This is a breathtakingly negligent way to determine students' grades. How could anyone ever consider that an algorithm could predict that three, and only three, students at 'Ambridge High' merited grade 4 for GCSE Geography? And likewise, across all subjects, across the country?

29. **The decision to use this algorithm was a grave error, and some questions the Inquiry might wish to pursue are:**

    - **Who took the decision to use the algorithm, and when?**
    - **What other possibilities were considered, and why were they rejected?** (In this context, Ofqual's 'Options Study', dated 16 March 2020, is particularly relevant (DHS/01) [**INQ000548401**])
    - **Who designed the algorithm and determined how it was to work?**
    - **Who wrote the code?**
    - **How was the algorithm tested?**

2020 – CAGs, rank orders and the Head's declaration

30. The cancellation of the summer 2020 exams in England was announced by the then Prime Minister Boris Johnson during his Downing Street briefing on 18 March 2020 (Prime Minister's Office, 2020), and, at essentially the same time, in a statement to the House of Commons given by the then Secretary of State for Education, (now Sir) Gavin Williamson. The Department for Education (DfE) issued a press release on 20 March (Department for Education, 2020a), but it was not until 3 April that Ofqual published two 'Guidance' documents describing what teachers and schools had to do (Ofqual, 2020a, Ofqual 2020b).

31. The two 'Guidance' documents comprise a total of 43 pages, but they boil down to a requirement on teachers and schools to send, to the appropriate exam board:

- For each individual student, and each subject, the 'Centre Assessed Grade', or 'CAG', this being the school's "*realistic judgement of the grade each student would have been most likely to get if they had taken their exams*".
- For each subject, a rank order of students.
- Overall, a declaration, signed by the school Head, confirming that the CAGs "*…honestly and fairly represent the grades that these students would have been likely to achieve if they had sat their exams as planned…*", with a similar affirmation as to the integrity of the rank orders.

32. That teachers should be invited to submit the CAGs appears to be very reasonable – teachers are being asked to use their judgement as to the grades their students deserve: as Gavin Williamson stated in the DfE's press release of 20 March, "*I have asked exam boards to work closely with the teachers who know their students best to ensure their hard work and dedication is rewarded and fairly recognised*" (Department for Education 2020a). As, however, I discuss in paragraphs 35 to 43, Ofqual's requirement on schools to submit CAGs was not only an enormous drain on teachers' time and emotional resilience ("Alex – grade 4 or 5? Oh dear, it's so hard to decide…"), but was also totally unnecessary. Furthermore, the presence of the CAGs alongside the results of the algorithm built in a delayed-action bombshell that was to explode as soon as the AS and A-level results were announced on 13 August (see paragraphs 145 to 159).

33. Even more of an emotional toll, however, resulted from the requirement on teachers to submit a rank order of students in each subject – a rank order that allowed only one

name on 'each rung of the rank order ladder', in that teachers were not permitted to make submissions of the form 'joint equals'. Teachers therefore had to split the finest of hairs between essentially equally-performing students, making judgements of which King Solomon himself would have been proud. This proved to be particularly difficult for teachers at Colleges of Further Education whose classes were largely of students re-sitting GCSE English or GCSE Maths, who all clustered around the grade 3/grade 4 boundary (Camden, 2020).

34. The final sting-in-the-tail was the Head's declaration – a declaration that put the entire weight of responsibility for the outcomes of the 2020 process on the shoulders of Heads and their staff, enabling the authorities – Ofqual, the DfE, the exam boards – to wash their hands of the grades actually awarded. This declaration put many Heads in an impossible dilemma, with repercussions that wreaked havoc after the algorithm was abandoned (see paragraphs 117 to 120).

2020 – Why the CAGs and the rank orders weren't needed

35. Right from the start, the underlying algorithm remained mysterious – so mysterious that not only was there no information as to how it would work, but, as noted in paragraph 27, it masqueraded under the term '*statistical standardisation*', described in Ofqual's documents of 3 April as "*a statistical model to standardise grades across centres in each subject* " such that, if a school's submitted CAGs "*...are more generous than expected, then the final grades for some of all of your students will be adjusted down; … if more severe…adjusted up*" (Ofqual, 2020a, page 11).

36. This is a most opaque way of describing the truth of what the algorithm actually did, which is best illustrated by a simplified example of an exam with only two grades, 'pass' and 'fail'.

37. Suppose that a school has 20 candidates, and submits CAGs of 'pass' for candidates ranked 1 to 15, and 'fail' for candidates ranked 16 to 20. If Ofqual's 'statistical standardisation' determines that only 9 passes are permitted, then the grades as awarded are 'pass' to candidates ranked 1 to 9, and 'fail' to candidates ranked 10 to 20.

38. Comparing the CAGs as submitted by the school to the awards actually made:

- Candidates 1 to 9 passed, in accordance with their CAGs.
- Candidates 16 to 20 failed, in accordance with their CAGs.
- Candidates 10, 11, 12, 13, 14 and 15 failed, with 'statistical standardisation' over-ruling the school's CAGs.

39. That explains why the school was asked to submit the rank order, and a rank order with 'only one person on each rung' – that's the only way that 'statistical standardisation' could 'slice' the list of students in the 'right' place, between candidates 9 and 10.

40. It is also apparent that the operation of 'statistical standardisation' is exactly the same as the childhood prank of "heads I win, tails you lose": if the CAG happened to be the same as 'standardisation', the CAG is confirmed - which is, of course, the same as 'standardisation' being confirmed; if the CAG is different from 'standardisation', then 'standardisation' over-rules the CAG. *Either way, the grade awarded is the result of the algorithm*.

41. This implies that the CAGs were not necessary, and that all the schools needed to have submitted was the rank order. But even the rank order was not required, for the result of the algorithm was to determine how many students were to be awarded each grade – in the example above, 9 passes and 11 fails. The simplest process would therefore have been for Ofqual to inform each school that, for each subject, the school was 'allowed' [this number] of [this grade], and asking the school to send the corresponding student names to the exam boards for the printing of the certificates, as discussed in my blog dated 18 June 2020 (Sherwood, 2020b).

42. **Given that an algorithm was used in 2020, by far the simplest process would have been to have told schools of their allocations accordingly, without the need for CAGs, rank orders or a declaration. This would have dramatically reduced the workload and the stress too.**

43. But by asking for CAG and rank orders, teachers were not only put under unnecessary pressure, but, as will be explained in paragraphs 150 to 158, Ofqual brought about disaster.

2020 – Misleading information

44. In the very first statement about the determination of the summer 2020 grades, the DfE's press release of 20 March, are the words *"The exam boards will be asking teachers, who know their students well, to submit their judgement about the grade that they believe the student would have received if exams had gone ahead"* (Department for Education, 2020a). This approach was confirmed in Ofqual's 'Guidance' of 3 April, which states that CAGs must "*reflect a fair, reasonable and carefully considered judgement of the most likely grade a student would have achieved if they had sat their exams this summer"* based on "*holistic professional judgement"*, as made by *"teachers and heads of department (who) will have a good understanding of their students' performance*" (Ofqual, 2020a, pages 4 and 5).

45. Anyone reading those words – and many other similar words too – is bound to infer that the exam boards would actually listen, and pay attention to, teachers' opinions. Especially given the obligation of every Head to sign the declaration that *"I am confident that (the submitted CAGs) honestly and fairly represent the grades that these students would have been most likely to achieve if they had sat their exams as planned"* (Ofqual, 2020a, pages 17 and 18).

46. I believe that those words are highly misleading, and were known to be misleading by those who wrote them. For the truth – a truth known to those writers – was that the CAGs were to play no role at all: the grades that were to be printed on candidates' certificates were always intended to be those generated by the algorithm, regardless of teachers' opinions and judgements. But not all teachers noticed that, for some believed that their judgement would carry weight. And some of those who did notice realised that agonising over whether 'Pat' truly merited GCSE Geography grade 5 or 6 was a total waste of time. Why not give CAGs of A* to all A-level candidates, and 9 for GCSEs, and let the algorithm sort it all out? That way, time can be spent where it really mattered – on deciding the rank order.

47. Only when the algorithm was scrapped did the CAGs have any significance. To the great surprise of the school that had submitted all A*s and 9s. And, no doubt, to the great delight of the students. In my view, the teachers at that school had not been negligent, nor had they 'gamed the system'. Rather, they had understood – and understood insightfully – exactly how 'the system' was intended to work. It was not the teachers' fault that 'the system' was fatally flawed, that the algorithm's results were scrapped, and

that – to everyone's great surprise – the CAGs suddenly had become real. But that all happened only after the A-level results had been announced, and only – in my opinion – because of Ofqual's fear that another part of 'the system', the appeals process, was about to collapse (see paragraphs 145 to 159).

2020 – The 'history trap'

48. On 15 April, Ofqual launched a 67-page consultation on their proposed process, announcing partial findings on 5 and 15 May, but it was not until 22 May that their final conclusions were published – and even then there were revisions as late as 16 June (Ofqual, 2020c). By mid-May, however, it became evident that the key purpose of 'statistical standardisation', as achieved by the still-mysterious algorithm, was to ensure that the distribution of the 2020 grades within any subject at any school corresponded to that school's 'history', as evidenced by that school's grade distributions over the preceding three years, 2017, 2018 and 2019 (Jadhav, 2020).

49. Although never stated by Ofqual, this was Ofqual's way of achieving its key objective of controlling grade inflation: if the distribution of grades for each subject at each school matches recent history, then the aggregate distribution of grades across all subjects at all schools will also match the recent past, and so grade inflation is suppressed.

50. This resulted in 'the history trap', whereby each school was a 'prisoner' of its history. Suppose, for example, that at a particular school over recent years, 5% of students had achieved the top grade, grade 9, in GCSE English; suppose further that in summer 2020, there were 20 GCSE English students. To comply with history, the number of students 'allowed' grade 9 would be 5% of 20, just one student. Suppose, however, that three of those students were 'Wiliam Shakespeare', 'Jane Austen' and 'Charles Dickens'. Clearly a very talented cohort. But only one can be awarded the top grade. Which one?

51. That example is of course theatrical. But it does make an important point. The cohorts in each year vary, and if 2020 happened to be more, rather than less, talented, the 2020 students were trapped by their less, rather than more, talented predecessors. A trap slammed tightly shut because Ofqual did not allow any opportunity for schools to offer evidence that all three of 'William', 'Jane' and 'Charles' truly deserved grade 9.

52. The school therefore faced a choice. To attempt to comply with their understanding of the 'rules', and give the grade 9 CAG to 'Jane', and grades 8s to 'William' and 'Charles',

despite their belief that all three truly merited grade 9. Or to ignore the 'rules', and give grade 9 CAGs to all three. If the school did that, however, there is also the requirement to determine a rank order – and if William and Charles are ranked 2 and 3, they will still end up with grade 8 after the algorithm has had its way. On top of all that, there's the declaration. The award of three grade 9 CAGs breaks the rules, but more "*honestly and fairly represents the grades these students would have achieved* "; one 9 and two 8s complies with the 'rules' and anticipates the expected outcome, but is problematic as regards the declaration… mmm… maybe William and Charles might have a 'bad day', so perhaps it's still OK to sign the declaration…

53. A variant of the 'history trap' concerns neighbouring schools. A particular school has a relatively weak history with no recent top grades, but a neighbouring school has a stronger history with several. Suppose further that, seven years ago, in 2013, an 11-year-old had a choice of school, and decided on the first school to stay with friends. That young person turned out to be very bright, but in 2020, could not receive a top A-level grade because the school had none 'available'. Yet had, seven years before, the choice been for the other school, that top grade would be quite possible. Such young people were truly trapped. Nor am I making that up. That is exactly what happened to Lexie Bell, as reported in *The Guardian* on 20 June (Lightfoot, 2020).

2020 – Different schools, different processes

54. Despite the fact that Ofqual's two 'Guidance' documents published on 3 April ran to 25 and 18 pages respectively, different schools adopted different processes to determine their CAGs and rank orders. To take just one example, the 'Guidance' states that, in determining the CAGs and rank orders, schools should "*draw on existing records and available evidence… including, where it is available… performance in mock exams"* (Ofqual, 2020a, page 6). In principle, that is sensible; in practice, some schools had carried out mock exams in a given subject, whilst others hadn't, and some schools had a policy of marking mock exams leniently (with the intention of being supportive to their students), whereas others marked stringently (with the intention of 'shocking' their students into better performance). Those schools with a more lenient policy for grading mocks were therefore in a position to submit higher CAGs than those who were 'tougher', even though the students might have been of equal ability.

55. Furthermore, some schools attempted to 'second guess' the algorithm in order to increase the likelihood that their CAGs will be confirmed, and not changed. This would

offer 'proof' that teachers can indeed be trusted, which would be a 'good thing' in general, and politically in particular. Accordingly, some of the teachers' unions, notably the Association of School and College Leaders (ASCL), gave their members detailed advice as to what, precisely, to do to help make this happen, and what records to keep as evidence of compliance with that process (ASCL, 2020), whilst technical experts, such as FFT Education Datalab (Thomson, 2020) and David Blow (Blow, 2020a) explained the intricacies of 'Comparable Outcomes' and 'Transition Matrices'. But even when attempting to follow the advice, some schools might give more weight to, say, homework, whilst others might have given more weight to classroom performance in the months before the lockdown.

56. Most teachers, without doubt, acted conscientiously and with integrity. But it was inevitable that a parent might think "if my child had gone to a different school, would the CAG have been different?". To which the answer is "quite possibly, yes". What is inevitable, however, is that the rank order position certainly would have been different, for the rank order is necessarily a comparison with others, and the 'others' in a different school are undoubtedly different.

57. **Had these very practical difficulties been sufficiently thought through by Ofqual before they adopted the 2020 process? I don't know. But perhaps the Inquiry might seek to find out.**

58. And talking of practical difficulties that had not been thought through, let me mention two more, these being technical problems associated with statistics, and with the need to round fractions and decimal numbers to whole numbers.

2020 – Problems with statistics

59. As discussed in paragraph 48, in essence, for each subject in each school, the process of 'statistical standardisation' determined the percentage of students assigned to each CAG by reference to the percentage of students awarded the corresponding grade in the preceding three years, 2017, 2018 and 2019 (except for GCSEs that had been numerically graded only since 2018, and for subjects 'new' to the school). If the CAG percentages complied with the historic pattern, the CAGs would be confirmed; if not, the pattern determined by the algorithm would over-rule the CAGs, and the rank order would be 'sliced' in the appropriate places to determine the grades to be awarded.

60. Those schools that wished to avoid many of their CAGs from being over-ruled therefore sought to ensure that the percentage distribution of their submitted CAGs matched the appropriate historic distribution as closely as possible.

61. When attempting to do this, schools very quickly discovered that, in many cases, the distributions of the grades awarded in any particular subject in each of the years 2017, 2018 and 2019 were different, especially for those subjects taken by a smaller number of students. What, then, was the 'target' distribution that the algorithm would be using as a 'match'? Is it the average over the three years? The highest achieved? The lowest? And suppose that a new teacher had been appointed 18 months ago, a teacher who had proven to be truly inspiring – so much so that everyone expected the 2020 exam results to break all school records. How might this be recognised?

62. Ofqual had no answers to any of these questions, and many like them. The rules of the algorithm were still undeclared. No school, no teacher, had any knowledge of what they were aiming for. Everyone was in the dark. So of course different schools took different decisions, resulting in the inevitability that the same student could well have been given a different CAG had he or she attended a different school – even though both schools were trying to be as conscientious as anyone might wish.

2020 – Rounding

63. This problem was much more prosaic. Rounding.

64. To illustrate, consider a school that, in each of 2017, 2018 and 2019, had entered 10 students for GCSE English, and, in each year, one student had been awarded each of the ten grades from 9 to 1 and U. There is no problem with different annual statistics: the distribution is rock-solid at 10% of students for each grade, as shown in Table 2:

Year	Number of students awarded each grade										Total number of candidates
	9	8	7	6	5	4	3	2	1	U	
2017	1	1	1	1	1	1	1	1	1	1	10
2018	1	1	1	1	1	1	1	1	1	1	10
2019	1	1	1	1	1	1	1	1	1	1	10
Distribution	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	100%

Table 2: Data relating to the number of students awarded each grade in each of 2017, 2018 and 2019

65. Suppose further that, in 2020, there are 9 candidates. Applying the historic 10% distribution results in 0.9 candidates for each of the ten grades. But '0.9 candidates' makes no sense – exam candidates come in whole numbers.

66. Of course. So let's round those numbers: 0.9 rounds to 1.0, giving one candidate for each of ten grades. But that implies that the total number of candidates is ten, as illustrated in Table 3. The reality, however, is that the total number of candidates is not ten, but nine. That 'extra candidate' is an artefact caused by the necessity of rounding to whole numbers.

	Number of students to be awarded each grade, 2020										Total number of candidates
	9	8	7	6	5	4	3	2	1	U	
Application of 10% historic distribution to a total cohort of 9 candidates	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	9
Rounding results in a total of 10 candidates, one too many	1	1	1	1	1	1	1	1	1	1	10
CAGs as submitted, for each of 9 candidates	1	1	1	1	1	1	1	1	1	0	9

Table 3: Applying the 10% historic distribution to the nine candidates in 2020

67. The solution to this problem is to allocate each of the nine candidates to nine of the grades, leaving one grade out. But which nine grades? Grades 9 to 1, with no-one assigned to grade U? Or grades 8 to 1, plus U, with no-one assigned to grade 9? Or some other allocation? Given that choice, what would you do? And would you be surprised if a school opted for grades 9 to 1?

68. Suppose further that there are 10 schools, A to J, each in exactly the same situation: each school, each with nine candidates, awards each candidate each of the grades 9 to 1. Those ten schools then send their CAGs to the exam board, resulting in the submission of ten CAGs for each of grades 9 to 1, and none for grade U – a total, as expected, of 90 CAGs, as expected for 90 candidates.

69. From the exam board's point-of-view, however, there are a total of 90 candidates and an associated grade distribution of 10% to each grade. That's nine candidates for each of the ten grades 9 to 1, plus U. But the number of 'bids' for each of the grades 9 to 1 is ten, with no 'bids' for grade U. That's a 'grade inflation' problem that needs to be 'controlled', as shown in Table 4:

	Number of candidates associated with each grade										Total number of candidates
	9	8	7	6	5	4	3	2	1	U	
Aggregate CAG 'bids' for 90 candidates from 10 schools	10	10	10	10	10	10	10	10	10	0	90
CAGs available as determined by applying a distribution of 10% to a total cohort of 90 candidates	9	9	9	9	9	9	9	9	9	9	90
Over (under) bids	1	1	1	1	1	1	1	1	1	(9)	

Table 4: 'Grade inflation' when each school's CAGs are aggregated

70. The exam board's solution is to downgrade one grade 9 candidate to grade 8, so awarding nine grade 9s, rather than ten. But which candidate is to be downgraded? The board has no reason to choose any one candidate over any other, so, since all schools have 'overbid' to the same extent, one school is chosen at random – say, school C – and the top-ranked student, whose CAG is grade 9, is downgraded to grade 8.

71. But there is a knock-on effect. For grade 8, there were originally ten 'bids', and now another – the downgrade from grade 9 – has been added, giving eleven 'bids' in total. But only nine grade 8s can be awarded, and so two grade 8s must be downgraded to grade 7. One of those is likely to be from school C, for – given the downgrade of the first-ranked candidate from grade 9 – there are now two grade 8s, but only one available. Accordingly, school C's first ranked candidate remains at grade 8, and the second-ranked candidate, originally given CAG grade 8, is downgraded to grade 7. And at the same time, at another school, say, school H, once again randomly chosen, the first-ranked candidate remains at grade 9, but the second-ranked candidate is downgraded from grade 8 to grade 7, leaving grade 8 void, and creating two grade 7s.

72. It doesn't stop there. For grade 7, there were ten original bids, and now two downgrades, giving a total of twelve, when only nine are available. This results in three downgrades from grade 7 to grade 6, the third-ranked students at each of schools C and H, and at a third school, A, too.

73. And so on across all the grades, ending up with nine downgrades from grade 1 to grade U, one at each of nine of the ten schools, as shown Table 5:

	CAGs as awarded following "statistical standardisation"										Total number of candidates
	9	8	7	6	5	4	3	2	1	U	
School A	1	1	0	1	1	1	1	1	1	1	9
School B	1	1	1	1	1	0	1	1	1	1	9
School C	0	1	1	1	1	1	1	1	1	1	9
School D	1	1	1	1	1	1	1	1	1	0	9
School E	1	1	1	0	1	1	1	1	1	1	9
School F	1	1	1	1	1	1	1	0	1	1	9
School G	1	1	1	1	0	1	1	1	1	1	9
School H	1	0	1	1	1	1	1	1	1	1	9
School I	1	1	1	1	1	1	1	1	0	1	9
School J	1	1	1	1	1	1	0	1	1	1	9
Total awards	9	9	9	9	9	9	9	9	9	9	90

Table 5: The results of 'statistical standardisation' for 10 schools, each submitting nine CAGs, one for each of grades 9 to 1. To control grade inflation, the exam board awards grades as shown. The shaded cells indicate the 45 downgrades, and only school D is awarded all the grades they 'bid' for. Note that the rank orders at each school are unchanged.

74. Although only nine grades were originally 'overbid' (one for each of grades 9 to 1), the strict imposition of a policy of 'no grade inflation', as executed by 'statistical standardisation', results – as shown in the Table 5 – in a total of 45 downgrades. Since there were only 90 candidates, that's 50% of candidates downgraded, including every candidate in School C, this being the school randomly chosen for the downgrade from grade 9 to grade 8. Only one school – school D – escapes unscathed. Why school D? No reason. Just luck.

75. No school had done anything wrong. No school had been malicious. No school had been 'playing games'. Each school had done its best to comply with its history. Each school had behaved both rationally and fairly given the need to round 0.9 to 1.

76. Yet the outcome is potentially catastrophic to the 45 downgraded candidates, none of whom has any idea as to why it happened to them.

77. And all attributable to Ofqual's failure to anticipate that fractions will have to be rounded, to offer explicit guidance as to how this should be done (and done in the same way at every school), and to take account of the consequences. Nor was this 'difficult' to anticipate, as verified by my piece in TES, dated 30 June 2020 (Sherwood, 2020c).

78. What negligence.

79. This example is of course an exceptional case. But it illustrates a very real point. Across the entire cohort of students taking any subject at any of the three English exam boards, if the number of 'bids' for top grade CAGs exceeded the algorithm's whole-cohort allocation, this would have a 'cascade' effect throughout all the grades, leading to many more downgrades than you might expect. As actually happened: on 13 August, when the algorithm's A-level grades were announced, nearly 40% of A-level CAGs had been downgraded (Adams, 2020a), as predicted on 23 July, using exactly this 'cascade' reasoning, by Huy Duong, the parent of an A-level student (Duong, 2020), as reported in the Guardian on 7 August (Adams, 2020b).

80. This problem – and the related technical problem concerning statistics, as discussed in paragraphs 59 to 62 – both have the same, easy, solution. A solution achieved by giving *every* school *the same, simple spreadsheet* requiring, as data for each subject, the number of students in this year's cohort, and also the school's historic results expressed not in percentages, but as the number of students awarded in each grade, for that's the information the school can easily access. All the rest – all the calculations, including the statistics and the rounding – is done within the spreadsheet, the output of which is a suggested distribution of grades that best fits the historic pattern. The school therefore does not have to worry about these technical matters, and the fact that the same spreadsheet is issued to all schools implies that *the same 'rules' are used in the same way* across the entire country, thereby ensuring 'standardisation'. The completed subject-specific spreadsheets from each school could then be submitted to the appropriate exam boards, where they could be checked and aggregated. It is, in fact, very easy to design such a spreadsheet – I myself did this in just a few days in mid-April (Sherwood, 2020d). That Ofqual failed to supply such a spreadsheet is woefully thoughtless.

81. Yes, there is a possibility that the overall result would show some grade inflation, primarily attributable to rounding up. Indeed. But is this really a 'problem'? Especially when such a process offers the important benefits of ease of use, and the confidence and trust associated with widespread knowledge that all schools, across the country, had used the same process, in the same way, applying identical 'rules' for technical matters such as the statistics and rounding?

82. Those were the main problems with the process that schools and teachers had to execute prior to submitting their CAGs and rank orders, as required between Monday 1 June and Friday 12 June; let me now turn to how events evolved during July and August.

2020 – Ofqual defies the Education Select Committee

83. During the early summer, there was increasing public concern about the absence of any information concerning how 'statistical standardisation' would actually work, especially as regards the possibility that the outcomes of the underlying algorithm might be biased with respect to particular groups, such as ethnic minorities or students with special educational needs – just one example being an important article written by the highly influential educationalist, Sir Jon Coles, published in Schools Week on 17 July (Coles, 2020); there are many others.

84. These matters were of particular interest to the House of Commons Education Select Committee, which, as early as 25 March, had initiated an inquiry into "*The impact of Covid-19 on education and children's services*". Many of the submissions are noteworthy in drawing attention to the multiplicity of problems associated with Ofqual's process, for example, those from Professors Lee Elliot Major and Stephen Machin (Major and Machin, 2020), and from The Royal Statistical Society (Royal Statistical Society, 2020).

85. The Committee's first report, "*Getting the grades they've earned - Covid-19: the cancellation of exams and 'calculated' grades*", published on 11 July (Commons Education Select Committee, 2020a), was highly critical, and of especial significance was recommendation number 5:

"Ofqual must be completely transparent about its standardisation model and publish the model immediately to allow time for scrutiny. In addition, Ofqual must publish an explanatory memorandum on decisions and assumptions made during the model's

development. This should include clearly setting out how it has ensured fairness for schools without 3 years of historic data, and for settings with small, variable cohorts."

86. There can be no doubt as to the clarity of those words.

87. Yet Ofqual refused to comply (Lough, 2020), delaying publication of the details to 13 August, alongside the announcement of the A-level results, by which time it was far too late for any beneficial remedial action to be taken. The damage caused by the algorithm's biases, approximations and inherent unfairness had been done.

88. Constitutionally, Ofqual is a non-ministerial department, independent of government, and reporting directly to Parliament (Ofqual, 2025a). Accordingly, Ofqual is not under the management or control of the DfE or the Secretary of State for Education, but accountable directly to the Houses of Commons and of Lords, within which the Education Select Committee is the senior body overseeing education.

89. How, then, is it possible for Ofqual to flout an instruction given by the Select Committee? And get away with it too?

90. **This is, I believe, a matter of great importance. To whom is Ofqual accountable? When there is a sense that actions being taken by Ofqual are heading in an inappropriate direction – as was indeed the case in the summer of 2020 – who has the power to intervene? And what sanctions can be invoked if Ofqual fails to comply?**

91. **These questions – which apply not only to Ofqual but to many other non-ministerial departments too (Ofsted comes immediately to mind), and are of significance beyond the events of 2020 and 2021 – are matters which I strongly hope the Inquiry will investigate.**

2020 – Gavin Williamson intervenes

92. As noted in paragraph 79, on 7 August, *The Guardian*'s front page, under the headline "*Nearly 40% of A-level result predictions to be downgraded in England*" (Adams, 2020b), described how Huy Duong, the parent of an A-level candidate, had put together some fragments of information, published by Ofqual on 21 July (Ofqual, 2020d), from which he had deduced that nearly 40% of A-level CAGs would be downgraded by the algorithm (Duong, 2020). The resulting publicity caused great consternation, significantly

enhancing anxiety that the algorithm would substantially contradict teachers' judgements, resulting in many lowered grades. How could an algorithm exercise wiser judgement than the teachers who, to quote Gavin Williamson once more, "*know their students well*" (Department for Education 2020a)?

93. *The Guardian*'s article was published four days after the Scottish exam results had been announced – Scotland's exam system is different from England's, with its own regulator, the SQA, but Scotland did adopt a similar process, asking schools to submit 'Scottish CAGs' which would then also be 'statistically standardised'. When the Scottish results were published on, 4 August, many were distraught to learn that, out of rather more than 500,000 grades awarded, some 133,000 had been over-ruled by the Scottish algorithm, 9,000 upwards, and 124,000 downwards (Scottish Government, 2020a).

94. This had caused uproar, so much so that, on 10 August, Scotland's then First Minister, Nicola Sturgeon, announced that there would be an "urgent review" of the downgraded results. That review happened very quickly, for on the next day, 11 August, the Scottish algorithm's results were scrapped, and candidates were told the grades on their certificates would be the higher of their (Scottish) CAGs and the results of the algorithm (Scottish Government, 2020b). This therefore set a precedent that the algorithm could be thrown away – a precedent based on the fact that 'only' some 25% of Scottish CAGs had been downgraded, in comparison to the prediction that 40% of A-level CAGs in England might be heading the same way.

95. Meanwhile, in England, CAGs and the algorithm filled the media, with talk of a "fiasco" and a "flawed educational system", highlighting that the workings of the algorithm were still unknown, that it might have any number of biases, and that it could not be trusted to deliver fair results.

96. The spotlight also fell on the appeals process – for, when 'normal' exams were held, if a candidate was unhappy with a result, there was the option to request a 'review of marking', offering the possibility of a change in the awarded grade if a 'marking error' is discovered and corrected. So if a candidate is unhappy with the results of the algorithm, surely the candidate could appeal too. And then everything would be alright.

97. What, then, was the appeals process for the grades awarded by the summer 2020 algorithm? That question was answered in Ofqual's 'Guidance' of 3 April (Ofqual, 2020, pages 16 and 17). Any candidate who was unhappy would have the opportunity to sit a 'real' exam in November 2020, with that result superseding the algorithm's, but no

appeal would be allowed against a result of the algorithm, a CAG, or a rank order position. These exclusions were justified by Ofqual on the grounds that:

- If the algorithm could be shown to have given the wrong result for one candidate, this would open the national floodgates for everyone else to appeal.
- If a CAG were shown to be wrong, that would raise doubts about all the other CAGs awarded by that school, opening the floodgates locally, and possibly beyond too.
- If the rank order position of any one candidate were to change, that would cause at least one other candidate's rank order position to change too, possibly changing that other candidate's grade, which would be unfair.

98. Ofqual's rules for 'appeals' not only, in essence, denied any appeal at all, but also stand as a tacit admission by Ofqual of the extreme frailty of the entire process, in that the discovery of an error – such as an individual candidate's position in the rank order – brings the whole 'house of cards' down. Furthermore, these rules were the single feature of Ofqual's consultation (see paragraph 48) that had attracted widespread disapproval (Ofqual, 2020e, pages 12 to 22, and Ofqual 2020f, pages 88 to 92). A disapproval Ofqual ignored. What arrogance. Again.

99. One consequence of the August pressure was therefore an announcement from Ofqual on 6 August confirming that neither the CAGs nor the rank order could be appealed, but that the results of the algorithm could be challenged under some specific circumstances, such as if a "monumental event" (for example, flooding or a fire) had occurred such that the historic results used by the algorithm might be unrepresentative, or if the "ability profile" of the 2020 cohort was likely to result in a "very different pattern of grades" as compared to previous years (Ofqual, 2020g). This was the first recognition by Ofqual that the algorithm might be flawed, but the new grounds for appeal were extremely limited, and left many concerns still unaddressed.

100. Then, on 12 August, the day before the AS and A-level results were due to be announced, the DfE issued a press release (Department for Education, 2020b) stating that the Secretary of State, Gavin Williamson, had decided, in order to "bolster fairness", that students could be awarded the highest result of this trio: the calculated grade (the result of the algorithm), the autumn exam grade (if the student opted to sit the 'real' exam in November 2020), or what was called "a valid mock grade". Gavin Williamson referred to this as the "triple lock". Notice, however, that the CAG is not included as a possible alternative, so even at this late stage, the CAGs were both ignored and irrelevant.

101. The reference to a "valid mock grade" was a great surprise, and puzzling too. What, precisely, is a "valid" mock as opposed to an "invalid" one? What about schools that did not have time to do mock exams before the lockdown? And what about the fact that schools do their mocks – and mark and grade them – in different ways: as already noted (paragraph 54), some schools mark mocks harshly, in the belief that the resulting 'shock' will encourage their students to 'pull their socks up', whereas others mark leniently in the belief that this will boost confidence.

102. Those details were clearly not a matter of concern to the officials of the DfE – as the press release of 12 August stated, "*Ofqual has been asked to determine how and when valid mocks can be used to calculate grades*".

103. If I might use a rugby metaphor… with his declaration of the "triple lock", fly-half Gavin Williamson has just thrown the mother-of-all-'hospital-passes' to centre three-quarter Sally Collier (Ofqual's then Chief Regulator), for, as I will discuss shortly (see paragraphs 135 to 144), this was a major contribution to the explosion that, five days later, on 17 August, was to blow the algorithm up.

104. **Did the DfE, or Gavin Williamson, consult Ofqual on the wisdom, or otherwise, of this intervention? I don't know. Perhaps the Inquiry might wish to find out.**

2020 – A-level results day, 13 August

105. On 13 August, the AS and A-level results were announced in England, Wales and Northern Ireland. Huy Duong's prediction that nearly 40% of A-level CAGs in England would be downgraded by the algorithm was proven to be correct (Adams, 2020a). The media exploded with indignation, whilst Ofqual's website posted a 'good news' bulletin proclaiming that "*96.4% of A-level awards are the same or within one grade of the CAG*" (Ofqual, 2020h). 96.4%! Wow! That's really good! Yes, 96.4% is an impressively big number, but those words "*same or within one grade of the CAG*" merit some thinking about… Does that mean that 96.3% of awards were the same as the corresponding CAGs, and that the algorithm had over-ruled only 0.1% of CAGs by one grade? And is that change upwards or downwards? Or have only 0.1% of CAGs been confirmed and 96.3% changed, either upwards or downwards? I can't tell… This ambiguity is yet another example of Ofqual's prowess in opaque communication.

106. The prize for communication, however, must go to the civil servant who drafted the blog posted on the DfE website, the following day, Friday 14 August (DHS/02) [**INQ000548402**]. Under the headline "*Misleading A-level claims debunked*", the text is the most explicit statement of the 'official' line, and exemplifies the arrogance that the authorities had showed ever since the cancellation of the exams in March. And, in the context of what was to happen just three days later, on Monday 17 August, when the results of the algorithm were scrapped, this blog is indeed an excruciating read.

107. Back in the real world, A-level candidates were dealing with the reality that nearly 40% of A-level CAGs had been downgraded – that reality, in numbers, being that, out of a total of 718,857 A-level grades awarded by the algorithm in England (Joint Council for Qualifications (JCQ), 2024), around 280,000 were downgraded from the corresponding CAGs. The total number of candidates was 274,685 (Ofqual, 2020i), just a few thousand below the number of downgrades. This comparison has an alarming implication: on average, *every candidate in England had a downgrade*. Every candidate.

108. And every candidate could make their own comparison, for once the 'official' results had been announced, schools were allowed to disclose a candidate's own CAG and rank order on request (Ofqual, 2020b, page 12). Every candidate could then compare their CAGs and rank orders with those of their friends, and wonder why the school had awarded that particular CAG and that particular rank order position. Why was 'Chris' higher in the rank order? Why had the algorithm over-ruled the judgement of *"the teachers who know their students best"* (to quote Gavin Williamson's statement of 20 March once more), particularly given that each Head had signed that declaration affirming that the CAGs *"honestly and fairly represent the grades that these students would have been likely to achieve"*? How could the algorithm know better? Especially since the algorithm had been much criticised ever since the Select Committee's report of 11 July, and the newspapers were now reporting all manner of 'strange' results, and alleged evidence of bias.

109. The following days were a spectacle of turmoil, with the media ablaze, students demonstrating outside Downing Street, and even talk of legal action (Elgot and Adams, 2020). And at an individual level, students and their parents were busy digging out those mock exam grades in preparation for an appeal…

110. …but, to return to the question posed in paragraph 101, what, precisely, was a "valid" mock grade? That ball had been left squarely in Ofqual's court, and on the morning of Saturday 15 August, Ofqual posted a 'news story' on their website describing their

answer (DHS/03) [INQ000548403]. But you had to be an early bird to read it, for anyone accessing that site later that day would find just these words: *"Earlier today we published information about mock exam results in appeals. This policy is being reviewed by the Ofqual Board and further information will be published in due course".* But that was never to happen…

2020 – Monday 17 August, the algorithm is scrapped

111. The uproar continued over the weekend, not just in England, but in Wales and Northern Ireland too. On Sunday 16 August, Northern Ireland announced that the GCSE grades, due to be published on Thursday 20 August, would be the CAGs, with the results of their GCSE algorithm discarded, but since the algorithm's A-level results had already been announced, those would stand (Department of Education., Northern Ireland, 2020). On the morning of Monday 17 August, the Welsh government announced that the forthcoming GCSE results would be the CAGs, and that the A-level awards, announced the previous Thursday, would be replaced by the higher of the CAG and the algorithm (Welsh government, 2020).

112. And at about 4 pm, Roger Taylor, the Chair of Ofqual, issued a statement that *"we have decided that students be awarded the centre assessment for this summer – that is, the grade their school or college estimated was the grade their school or college was the grade they would most likely have achieved in their exam – or the moderated grade, whichever is higher"* (Ofqual, 2020j). This applied in England, both to the GCSEs to be announced the following Thursday, and also retrospectively to the A-level and AS grades announced on 13 August.

113. Shortly thereafter, Northern Ireland declared that their A-levels too would be the higher of the CAG and the algorithm (CCSA, 2020). By the late afternoon on that Monday, all the algorithms used across the four UK nations had been blown up.

2020 – The aftermath, and the blame

114. Tuesday 18 August 2020 is probably the only day in history on which every national newspaper – from the *Times* to the *Daily Star* – carried the same lead story on their front pages, and with that story being about exam grades. To choose just a few extracts from the headlines… *"U turn*". "*Fiasco*". "*Another fine mess*". "*A-level and GCSE marks torn up after outrage*". "*Why is Education Secretary still in job*?" (National newspapers, 2020).

115. In fact, Gavin Williamson remained as Secretary of State for Education until September 2021, and was rewarded with a knighthood in March 2022; as already mentioned (see paragraph 18), the only people who 'moved jobs' were Ofqual's Chief Regulator, Sally Collier, and Jonathan Slater, the Permanent Secretary at the DfE.

116. **The Inquiry may wish to pursue the important matter of accountability.**

117. And rather than holding those truly responsible for the "fiasco" to account, much of the blame ended up – to my mind most unfairly – on many individual teachers, and on the teaching profession collectively. For now the students' grades were the CAGs, and the CAGs had been determined by teachers and the schools. And hadn't the Head signed that declaration confirming that *"these centre assessment grades, and the rank order of students … are accurate and represent the professional judgements made by my staff…"*?

118. So if there was suspicion that students at private, fee-paying schools had generally received higher grades than students at state-funded schools, that's because private school teachers had conspired to 'game the system'. If there was suspicion that special needs children received lower grades, that's because their teachers not only don't understand them, it's because they're biased too. If 'Kim's' parent believed that the grade should have been A* rather than A, surely that was the result of the vendetta 'Dr Hardman' had been raging ever since that parent had complained about the mark for that piece of homework 'Kim' had done in October 2018…

119. Needless to say, Ofqual, the DfE and the exam boards were nowhere to be seen. "Problem with the CAGs? Nothing to do with us…".

120. As noted in paragraph 97, long before the algorithm was scrapped, Ofqual had ruled that CAGs and rank order positions could not be appealed. But that did not stop all sorts of allegations in the media, or that determined parent – a parent who, on being informed by the school "But you cannot appeal against the CAG", had no hesitation in expressing an opinion. But perhaps an opinion that 'Kim' might have preferred not to be expressed, for maybe what 'Kim' really wanted was to get on with things, rather than an unending, and increasingly acrimonious, dispute with the school, and the school's teachers.

121. Many of the issues concerning the possibilities of bias and 'game playing' could have been resolved from study of the CAG data submitted by schools. If, for example, a school

had submitted A*s for all their A-level students, the evidence would be visible in those submissions. It is understandable that a school would not wish to make this information available to any one aggrieved student or parent, for the relevant data relates to whole cohorts, and each data item is of course personal to each individual. But study by, for example, an academic who has no vested interests, and who would honour the need for confidentiality, is very different. The suggestion that this data should be made available for detailed impartial scrutiny was, I believe, first made as early as 12 August, before the AS and A-level results were announced (Sherwood, 2020d); furthermore, in reply to a question about this at the hearing of the Education Select Committee on 2 September 2020, Ofqual's then Chair, Roger Taylor, stated that "*It is absolutely essential that independent researchers have access to that (data) in a secure way that will enable those lessons to be learned*", and committed to do just that (Commons Education Select Committee, 2020b). Yet it was not until 29 April 2022 that the data was actually made available (Ofqual, 2022).

122. **Why did it take Ofqual some 20 months to comply with the Select Committee's request? Is this another example of Ofqual's apparent immunity from accountability?**

2020 – My perspective

123. Let me now take this opportunity to step back, take a wider perspective, and offer my opinions as to what was happening and why; may I also note that there are many other perspectives too, of which those of David Blow are particularly insightful and authoritative (Blow, 2020b).

124. Most importantly, the process had been explicitly designed so that the CAGs would play no role whatsoever: as discussed in paragraphs 35 to 40, the algorithm would always 'win'. That raises the question as to why they were asked for in the first place, a matter I will explore further in paragraphs 150 to 159. Given, however, that the CAGs were required, those who realised that they did not matter could well have submitted, say, all A*s or 9s. But even those who had been conscientious had expected that the grades shown on candidates' certificates would be those generated by the algorithm. So it was a great surprise to everyone – including the author of the DfE blog posted on the day after the algorithm's AS and A-level results were announced (see paragraph 106) – when the algorithm was discarded, and when the CAGs suddenly not only had a role, but a vital one.

125. Despite the fact that, in principle, the CAGs were of no consequence, many schools were extremely diligent in determining them, holding meetings with the appropriate teachers, scrutinising homework marks, keeping detailed records, striving to be fair. Many also wished to comply with what they understood to be 'the rules', and, in particular, with 'statistical standardisation'. Those 'rules', though, were not divulged by Ofqual until A-level results day. Schools were therefore obliged to make guesses, aided by guidelines issued by, for example, ASCL (ASCL, 2020), or with support from expert educational statisticians, such as the team at FFT Education Datalab.

126. Overshadowing everything, however, was the knowledge that Ofqual's overarching objective during most of the last decade had been 'to control grade inflation', and that their means of achieving that objective in 2020 was to use 'statistical standardisation' to ensure that the 2020 grades had the same distribution as in recent years. In practice, this caused technical problems such as those associated with variations in the historic distributions (see paragraphs 59 to 62), the need for rounding fractions and decimals to whole numbers (see paragraphs 63 to 82), and also the profound ethical problem of the 'history trap' (paragraphs 48 to 53).

127. As noted in paragraphs 50 to 52, the Head of the school where 'William Shakespeare', 'Jane Austen' and 'Charles Dickens' were students faced a difficult dilemma. According to the school's history, only one A-level grade A* was 'available'. So if the Head wished to comply with the 'rules' – or felt obliged to by virtue of the guidance being followed at the school – then a choice had to be made as to which one candidate would be given a CAG of A*, and therefore ranked number 1, with the others as numbers 2 and 3. Alternatively, if the Head wished to recognise the candidate's true ability, then all three would be given A* CAGs. That decision, however, was taken in June, at which time it was certain that the algorithm would downgrade two of the three A*s, so ending up in the same place as the first option.

128. But not quite the same place. Although the students will be awarded A*, A and A either way, from the school's point-of-view, there are some important differences. The option of giving CAGs of A*, A, A shows that the school 'followed the rules', whereas if the CAGs were A*, A*, A* those rules were not only broken, but – should the CAGs ever be disclosed – might lead to a general perception that the school had 'gamed the system', for it would be very difficult to get involved in a public argument attempting to 'prove' that all three of 'Will', 'Jane' and 'Charlie' are really very talented. That makes the first option

– one A* – the safer. But for one issue. The declaration. The declaration is only 'true' under the second option, three A*s.

129. So should the Head sign a 'false' declaration and submit A*, A, A – with a very high likelihood that those CAGs will all be confirmed – and so be seen to have complied with the 'rules'? Or sign a 'true' declaration, submit three A*s, get awards of A*, A, A, anyway, and risk being accused of ignoring the 'rules', or – worse – of 'playing games'?

130. A further, real, complication is attributable to the fact that the CAGs were likely to be determined, in the first instance, by the subject teacher, or, for larger cohorts, a number of different teachers, each of whom had taught different students. If 'Will', 'Jane' and 'Charlie' are in different sets, the first confrontation is when the three respective teachers have to agree, or perhaps fight over, which – or rather whose – student gets the A*. But suppose that the three English teachers agree that the 'right' answer is three A*s. That recommendation then goes to the school's leadership team, none of whom know 'Will', 'Jane' or 'Charlie' other than by name. The leadership team's priorities are different, and they are far more alert to the politics than are the subject teachers. Not just that. At the meeting to review all the teachers' recommendations, they suspect that the English Department is 'trying it on'. So they over-rule the English teachers and choose 'Jane' for the A*. As actually happened, as described by an 'anonymous' teacher in an article in The i Paper (The i paper, 2020).

131. Every school faced any number of these dilemmas, for the process of determining the CAGs and rank orders was highly stressful.

132. And when, on 13 August, the algorithm was discarded and the CAGs reinstated, everyone was surprised. Really surprised.

133. Had that school submitted A*, A, A, then, within minutes, the highly irate parents of 'Will' and 'Charlie' would have been on the phone; had the school submitted three A*s, then it was 'Madame Dumas' ringing through, alleging that the grade A awarded to 'Alexandre' had been biased because his first language was French.

134. Furthermore, if a school, seeking to comply with 'the rules', had in fact awarded CAGs that were now recognised as being lower than then grades the school believed the students truly deserved, any hopes that the school might appeal and get higher grades awarded were soon dashed (Schools Week, 2020). The CAGs would remain unchallengeable, even if they were acknowledged by the school to be wrong.

2020 – Why was the algorithm discarded?

135. Continuing with my own opinions… let me answer my own rhetorical question with my own view. My hunch is that some time on the afternoon of Saturday 15 August, or morning of Sunday 16 August, the Board of Ofqual decided to abandon the algorithm for fear that the appeals process would break. That's because Gavin Williamson's "triple lock" intervention, declaring that "valid mocks" would constitute grounds for appeal, was impossible to implement robustly and fairly, and – even if it were – the appeals system would be overwhelmed, not just by the AS and A-level appeals, but by the likely number of GCSE appeals following the announcement of the algorithm's GCSE results the following Thursday.

136. Let me stress that I made that up. I wasn't at the Board meeting. So let me explain…

137. Most importantly, Ofqual had staked all on the algorithm, even to the extent of defying the Select Committee (see paragraphs 83 to 91). As far as Ofqual was concerned, when the A-level results were announced on 13 August, the algorithm had done its job, as the DfE bulletin of 14 August makes vividly clear (DHS/02) [**INQ000548402**]. Certainly, there would be grumbles and criticisms, especially once the details of the algorithm had been published (as they were on 13 August), laying bare its weaknesses. But those storms could be weathered…

138. Furthermore, the possibility that there might be disruption attributable to appeals had been 'managed' by the 'rules'. The original rules, as confirmed by Ofqual on 27 July, denied any challenge to the results of the algorithm, the CAGs, and the rank order, so all the key elements of the process were 'protected' (Ofqual, 2020a, pages 16 and 17). Appeals would be allowed, however, on grounds such as malpractice, maladministration, bias or discrimination, but those would be disputes between a student and the school, and so of no concern to Ofqual, the DfE or the exam boards. Later, on 6 August, and then only after much public pressure, did Ofqual broaden the criteria for appeal to include, for example, a "monumental event" (see paragraph 99). Those additional grounds would indeed increase the number of appeals that would need to be processed, but not unduly.

139. That all changed with Gavin Williamson's declaration concerning "valid mocks". From Ofqual's standpoint, a "valid mock" would, in practice, be very hard to distinguish from

"the test done that Friday last November". Ofqual therefore faced the possibility, that, with not too much ingenuity, every school would attempt to produce evidence that "that test" was indeed "valid", and the volumes of appeals would be overwhelming – remember that some 280,000 CAGs had been downgraded, an average of one for every A-level candidate in the land, and if they were all to be appealed, with the great majority needing to be resolved in time for university admissions…

140. Even worse was the prospect of what might happen when the 5.2 million GCSE results were to be announced on Thursday 20 August. Given that the algorithm had been scrapped by then, there is no published information about how many GCSE CAGs were downgraded. But over that weekend, Ofqual did know. Suppose that the percentage for GCSE was about the same as it was for A-level, around 40%. That's more than 2 million downgrades. And possibly 2 million appeals based on "valid mocks". The appeals system would most surely break.

141. So, at that Board meeting, what were the possible options?

142. One might have been to go back to Gavin Williamson, and say, "Actually, Secretary of State, perhaps it might be appropriate to amend somewhat the possibility of using "valid mocks" as grounds for appeal…". Perhaps the Board discussed that; perhaps that happened. But if it did, Gavin Williamson must have said "no".

143. A second might have been to press on regardless, reinstating the guidelines that had been published on the Ofqual website on the Saturday morning, and hope for the best…

144. And a third was the unthinkable. To avoid the appeals catastrophe by throwing the algorithm away and reverting to the CAGs. After all, there was the precedent in Scotland. Thank goodness that we asked for the CAGs, for surely the CAGs are the back-stop. And a 'safe' backstop too. Of course there will be some flak for abandoning the algorithm, but as long as someone steps down (eyes go round the table) that will all blow over. And if students don't like the CAGs, that's not our problem…

2020 – Why things really went wrong

145. My (most hypothetical) re-enactment of the Ofqual Board Meeting imagines the relief that might have been felt when it was realised that the CAGs were the back-stop. Yes, they were. But a far deeper truth is that the CAGs were the fundamental problem.

146. To explain, let me go back to a 'normal' year, say, 2019. On A-level results day, 'Ali' is informed of the award of grade B for A-level Psychology. 'Ali' may be pleased; 'Ali' may be disappointed. But how does 'Ali' know whether or not that grade B is right?

147. 'Ali' doesn't.

148. That grade B has to be taken on trust, for there is no comparator of 'right' against which the grade B can be tested. The only possibility, if 'Ali' is unhappy, is to appeal, and see what happens then.

149. This is a fundamental feature of the entire exam system. It all relies on trust – trust in the single grade that is awarded.

150. Summer 2020 was very different. That's because, for every grade determined by the algorithm, *every student in the land had a comparator – the CAG*. And if the CAG was different from – and, in every likelihood, higher than – the grade determined by the algorithm, the student could ask "*why?*".

151. This comparison was, of course, central to the concept of the "downgrade". A "downgrade" is inherently relative – a difference between one grade and another, the difference between the CAG and the result of the algorithm. With the psychological overtone, as implied by 'down', that the CAG is somehow more 'true' than the algorithm.

152. Suppose, then, that the CAGs (and rank orders too) had not been demanded from schools, and that Ofqual had – as described in paragraph 41 – told each school "your allocation of grades for (this exam) is shown here – please complete the attached form with the names of the candidates to be awarded the specified number of the corresponding grades".

153. The schools would have complied, and returned the forms, duly completed with students' names. There may have been – in fact undoubtedly would have been – any number of complaints along the lines of "but how can a machine know how many grade 5s…?". But those clever people at Ofqual would have had cleverly crafted answers, as did the author of the DfE blog of 14 August (DHS/02) [**INQ000548402**], so they would have survived that.

154. On 13 August, the results would have come out, and students informed of their grades. Just like a 'normal' year. Some students would have been pleased; some not. Just like

a ‘normal’ year. And those that were unhappy would complain. Just like a ‘normal’ year. But unlike a ‘normal’ year, that complaint would be aimed at the school. Not at Ofqual. Not at the exam board. Well, that’s the school’s problem…

155. The **BIG DIFFERENCE** would have been that, with no CAGs, there is no possibility of headlines such as “*40% of teacher predictions to be downgraded*”, and no possibility that any student could ask “why has the algorithm given a result lower than my teacher thinks I deserve – and signed a declaration to that effect too?”.

156. Ofqual’s **BIG MISTAKE** was to ask for the CAGs. By doing so, Ofqual *built in an appeals process accessible to every student*, enabling *every student* to compare one judgement (the school’s CAG) with another (the results of the algorithm), with – in every student’s eyes – the CAG having far more legitimacy and credibility than the algorithm, especially if the CAG was higher (Sherwood, 2020e).

157. Yes, the appeals process did indeed bring the whole edifice down. But it wasn’t the post-results appeals process that failed. It was the implicit appeals process inherent in the requirement for schools to submit CAGs, as explicitly and deliberately specified by Ofqual in March and April. Ofqual really hadn’t thought things through – let alone done what they should have done, which was to rely on teacher judgement, and work with the school unions to ensure that the process had integrity. Ofqual were the architects of their own disaster.

158. None of the ideas in the last few paragraphs are ‘rocket science’; none are ‘with the benefit of hindsight’. All could have been thought through in advance, at the outset, in March 2020. Yet Ofqual decided to use an algorithm, and to require the schools – quite unnecessarily – to submit CAGs and rank orders, with catastrophic, and long-lived, consequences. Why did they do that? Was it because they believed that they needed to be seen to ‘involve’ teachers, even if that ‘involvement’ was both burdensome and a sham? Or was it because they feared that the algorithm would never work, and that they needed a back-up, the CAGs, just in case the worse happened…?

159. I don’t know. **Perhaps the Inquiry might wish to find out.**

2021 – The TAGs

160. In the autumn of 2020, everyone hoped – and indeed expected – that Covid-19 was in the past, and that every-day life would return to pre-Covid conditions. But, as we know, that was not to be. On 7 October, Scotland – so often one step ahead of England – announced the cancellation of their summer 2021 National 5 exams (the Scottish equivalent of GCSE), and that students' grades would be determined by teacher judgement, without the 'aid' of an algorithm; their higher level exams were cancelled on 8 December. Wales cancelled all their exams on 10 November; England cancelled on 4 January 2021; Northern Ireland on 6 January 2021.

161. The process to be used in England was announced by Gavin Williamson, still in post as Secretary of State for Education, in a speech to parliament on 6 January: "*This year, we are going to trust in teachers rather than algorithms … (using) a form of teacher assessed grades with training and support provided to ensure that these are awarded fairly and consistently across the country"* (Williamson, 2021).

162. These words, however, were preceded by "*although exams are the fairest way of assessing what a student knows…"*, so setting up the conundrum:

- Exams are the fairest way.
- Teacher assessments are not exams.
- Teacher assessments must therefore be less fair than exams.
- But we're going ahead with teacher assessments anyway.

163. Gavin Williamson was under no obligation to preface "*we are going to trust teachers"* with *"although exams are the fairest way"*. But in choosing to do so, is he revealing the official mind-set that the authorities really don't trust teachers and never have (see paragraphs 22 to 24)? And in so doing, did this undermine the 2021 process?

164. And with Ofqual, the DfE and exam boards nowhere to be seen, teachers were on their own. Were they being set up to fail? For if teachers could be discredited even more than following the 2020 CAGs, then any trust that might still be attached to teacher judgement would be shot to pieces, strengthening the position of the "*exams are the fairest way*" brigade in any discussions about how student assessment might evolve in the future, once Covid-19 really was in the past (Department of Education, 2020c). It is surely no accident that the 2021 process was associated with 'TAGs' – *teacher* assessed grades

– in contrast to the 2020 'CAGs' – *centre* assessed grades – emphasising that the awards made in 2021 were those determined not by a nebulous 'centre' but by individual teachers. And only teachers. Alone.

165. The 2021 process, confirmed by Ofqual on 25 February, was very much simpler than in 2020. For each student, teachers were to submit their recommended TAG, which would be reviewed by the appropriate exam board as being a "*reasonable exercise of academic judgement".* No 'statistical moderation'. No requirement to match any particular historical pattern. Just teacher judgement, subject to the exam board confirming that the submitted TAG is not "unreasonable" (Ofqual, 2021a, pages 11-13), As regards appeals, any concerned candidate could challenge the school in the first instance, alleging that a TAG was indeed "unreasonable", and, if unhappy with the resulting outcome, could ask the exam board to check the grade for "reasonableness".

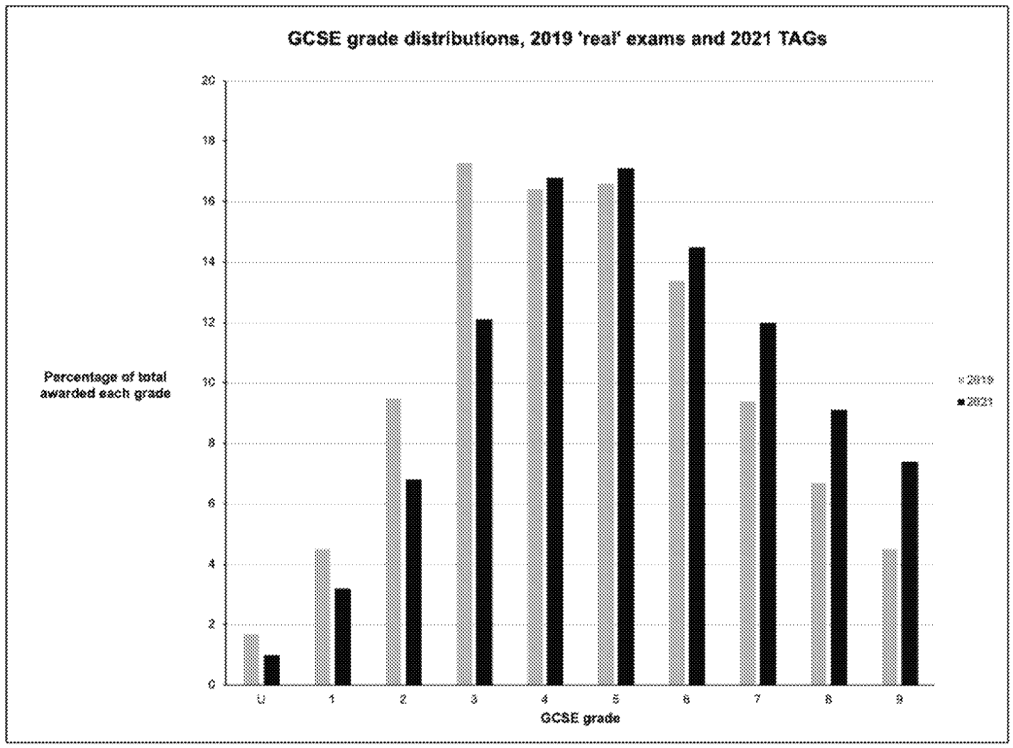
166. The 2021 process was therefore far less burdensome for teachers, who could exercise their judgement as they wished. One important difference as regards 2020, however, was the fact that teachers had less information on which to draw: the 2020 cohort had started their GCSE and A-level courses in September 2018, and so had nearly a year-and-a-half of school tuition before the first lockdown; the 2021 cohort, having started in September 2019, had spent far less time in the classroom, with continuous teaching re-starting only when schools reconvened on 8 March 2021. It might have been possible for the authorities, that March, to have thought, "now everyone's back at school, let's re-instate the normal exam process for the coming summer". Perhaps someone did think that. But after the 2020 fiasco, the authorities were keeping the lowest of profiles, so grading would still be by TAG. Which probably had the benefit that school time as of March 2021 could be devoted totally to catching up on lost learning, without the 'distractions' of mock exams, revision and the exams themselves.

167. As regards the TAGs, the key requirement was not to be "unreasonable". For those A-level students applying to university, a possibility was already on the table – the 'UCAS prediction', as required to support each candidate's application. The UCAS guidelines require this prediction to be "*aspirational but achievable*" (UCAS, 2025), which many people would consider to imply "reasonable" too. So if a student's TAGs were the same as the UCAS predictions, it would be very difficult for an exam board to consider them to be "unreasonable", and the student is likely to be happy too.

168. For GCSEs, there is no equivalent to A-level UCAS predictions, but since – understandably – teachers want to do the best for their students, the GCSE TAGs would

surely give students ‘the benefit of the doubt’, especially at that critical 3/4 grade boundary for GCSE English and Maths.

169. So when the results were announced in August 2021 (Ofqual, 2021b), no one was surprised that the percentages of top grades were higher than they had ever been, as shown by the numbers in Table 1 (see paragraph 6), and visually in the following Figures, which compare the GCSE and A-level grade distributions for the ‘real’ exams of 2019 and the TAGs of 2021:



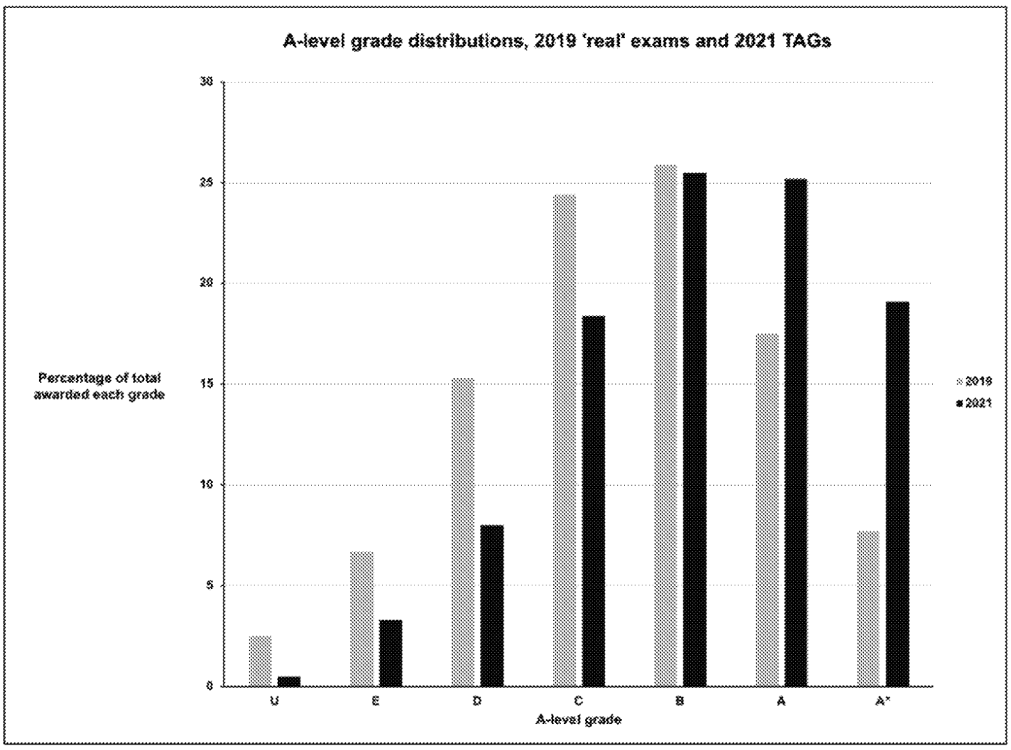


Figure 1: Grade distributions, 2019 and 2021's TAGs

170. I'll return to these histograms in paragraphs 217 to 234; for the moment let me draw attention to the shapes of the GCSE distribution, and in particular to the height of the grade 3 column for the 2019 'real' exam grades (in grey, to the left). With over 5 million awards distributed across the ten grades, my expectation would be that the resulting overall shape would closely resemble what mathematicians and statisticians refer to as a 'Gaussian' or 'normal' distribution, with its characteristic 'bell' shape. This the case for the 2019 A-level exam distribution, as well as for the two TAG distributions, albeit that both TAG distributions are somewhat shifted to the right, towards higher grades.

171. To me, the 2019 GCSE 'real' exam distribution is somewhat odd, especially as regards grades 3 and 4: if the distribution were Gaussian, the column corresponding to grade 3 would be lower, and that for grade 4 higher. It's as if there are too many grade 3s and too few grade 4s. Does that raise questions as regards Ofqual's setting of the grade 3/grade 4 boundary? Had, for example, that boundary been set, say, one or two marks lower, then there would have been fewer grade 3s and more grade 4s, resulting in a shape that was much closer to the 'normal' distribution – as is the shape of the TAGs. Might this suggest that Ofqual had set the grade 3/grade 4 boundary harshly high?

172. If the lockdowns had not happened, Ofqual would have continued its policy of 'no grade inflation' in both 2020 and 2021, enforcing grade distributions comparable to those of the late 2010s, of which the 2019 distribution is representative. Had that distribution been applied to the 2021 cohort, the numbers of each grade awarded would have been as shown in the following table (Joint Council for Qualifications (JCQ), 2024):

GCSE grade	Grade distribution, %		Number of awards		TAGs over / (under) 2019 'real'	
	2019 'real' exams	2021 TAGs	2021 TAGs	2021 as 2019	Number of awards	% relative to 2019
U	1.7	1.0	52,369	89,027	(36,658)	(41.2)
1	4.5	3.2	167,580	235,659	(68,079)	(28.9)
2	9.5	6.8	356,107	497,502	(141,395)	(28.4)
3	17.3	12.1	633,660	905,977	(272,317)	(30.1)
4	16.4	16.8	879,792	858,845	20,947	2.4
5	16.6	17.1	895,503	869,318	26,185	3.0
6	13.4	14.5	759,345	701,739	57,606	8.2
7	9.4	12.0	628,423	492,265	136,158	27.7
8	6.7	9.1	476,554	350,870	125,684	35.8
9	4.5	7.4	387,528	235,659	151,869	64.4
Total	100.0	100.0	5,236,861	5,236,861		

A-level grade	Grade distribution, %		Number of awards		TAGs over / (under) 2019 'real'	
	2019 'real' exams	2021 TAGs	2021 TAGs	2021 as 2019	Number of awards	% relative to 2019
U	2.5	0.5	3,763	18,814	(15,051)	(80.0)
E	6.7	3.3	24,834	50,421	(25,587)	(50.7)
D	15.3	8.0	60,204	115,141	(54,937)	(47.7)
C	24.4	18.4	138,470	183,623	(45,153)	(24.6)
B	25.9	25.5	191,901	194,911	(3,010)	(1.5)
A	17.5	25.2	189,644	131,697	57,947	44.0
A*	7.7	19.1	143,738	57,947	85,791	148.1
Total	100.0	100.0	752,554	752,254		

Table 6: Applying the 2019 grade distributions to the 2021 cohort

173. In these tables, the left-hand columns show the percentages of awards corresponding to each grade for the 2019 'real' exams and the 2021 TAGs. In the centre, under the heading 'number of awards', the column '2021 TAGs' shows the actual number of 2021 TAG awards for each grade, and the adjacent column, '2021 as 2019', shows the number of awards for each grade as would have been made had the grade distribution been that of the 'real' exams in 2019. To the right, the two columns under the heading 'TAGs over/(under) 2019 'real' show two comparisons of the numbers of TAGs as awarded for each grade and the corresponding number using the 2019 distribution. So, for example, the number of A level grade A TAGs actually awarded was 189,644; had the grade distribution been that of 2019, the number would have been 131,697. Accordingly, in 2021, there were 189,644 – 131,697 = 57,947 more grade As awarded than there would have been had Ofqual adopted a strict policy of 'no grade inflation'. Also, the number of TAG grades awarded, 189,644, is 44.0% greater than the 'no grade inflation' number of 131,697.

174. The key message from Figure 1 and Table 6 is that the TAGs resulted in a significant increase in top grades, and a correspondingly significant decrease in the lower grades, so providing ample ammunition to those who argued that "You can't trust teachers – they always reward their students too highly. That's why we need exams. After all, exams are the fairest way, and the 2021 TAGs are the proof". Whether exams are indeed "the fairest way" is indeed an important issue, which I discuss further in paragraphs 179 to 202, and 217 to 234.

175. As regard the impact on the students, the higher number of higher grades suggests that many were delighted, and their parents too. And, unlike in 2020, when every student could compare the results of the algorithm with their CAGs and ask "why have I been downgraded?", in 2021, there were no corresponding 'rival' grades: students were awarded their TAGs, just as, in 'normal' years, students are awarded their exam grades, and that's that.

176. Some students were undoubtedly unhappy with their TAGs, but there was far less anguish than in 2020. Those that were unhappy were then mired in disputes with their schools as to what was, and was not, "reasonable" – disputes in which students (or more likely, their parents) were energised by the general "you can't trust teachers" feeling. In reality, though, the dice were heavily loaded in favour of the school, not least because their TAGs had, ostensibly, been reviewed by the exam boards when they were submitted, and therefore already affirmed as "not being unreasonable".

177. Although students rejoice in higher grades, let me note that being awarded a grade higher than might be truly merited might bring some problems too. If, for example, the student enters a course pitched at a high level, that student – although qualified to participate according to the grades awarded – might struggle, suffer undue mental health problems, sense 'imposter syndrome', perhaps drop out. Might that have happened to some of the summer 2021 cohort? Quite possibly; I don't know. But perhaps some data might be available from, for example, universities. Has the drop-out rate associated with the October 2021 entry been unusually high? And is there a correlation with the A-level TAGs?

178. One other aspect of the 2021 cohort – for both GCSE and A-level students – relates not to the TAGs but to the loss of learning attributable to missing so much 'normal' schooling. Those taking their GCSEs and A-levels in 2021 inevitably had learnt less 'stuff' than their counterparts in 2018 and 2019. Did subsequent A-level programmes or university courses have to offer 'remedial' instruction to 'fill the gap'? Or were students able to cope? The former would be understandable; but if the latter, what does that imply as regards the GCSE and A-level curricula?

Are exams the "fairest way"?

179. As noted in paragraph 166, schools reconvened in March 2021, and continued 'normally' thereafter. And, in accordance with the policy that "*exams are the fairest way*" (Willamson, 2020), sit-down GCSEs and A-level exams returned in summer 2022, and have remained unchanged ever since. Furthermore, on 30 September 2021, Ofqual's then recently-appointed Chief Regulator, Do Jo Saxton, announced that, in accordance with the policy of 'no grade inflation', grade distributions in summer 2023 would return to those of the late 2010s, with summer 2022 being a "transition year" in which the grade distributions would be set mid-way between those of the TAGs of 2021 and the last year of 'real' exams, 2019 (Saxton, 2021). As indeed has happened.

180. But are "exams the fairest way"? And if so, what does that actually mean?

181. There are, in fact, many (very well-known) reasons why exams are not fair at all. Is it "fair" that 16 year-old students, who have had 11 years of development and attendance at school, are judged by their performance in a single written GCSE exam, taken on a single hot summer's day? And have their entire school experience recognised by a

single letter or number grade? By comparison, over the last many months there has been widespread outrage at the 'single word' judgements on schools given by the school inspectorate, Ofsted, on the basis of an on-site inspection of just a few days (Crerar, 2024). If a 'single word' is inappropriate for the judgement of a school, how can a single symbol be appropriate for a student?

182. Furthermore, we all know that the same student, taking the same exam paper, but on a different day, could well get a different mark, possibly resulting in a different grade. Perhaps, on a particular day, the student might have been unwell, tired, under undue stress, or suffering from some form of anguish, all of which suggest the student is likely to under-perform. On another day, the student's personal circumstances could be very different. Similarly, if, on any particular day, the student were to sit a different exam paper, with different questions – but ostensibly of the same standard as the 'first' paper – then it is also quite likely that the student would get a different mark, and perhaps grade. We all know about being 'lucky' – or indeed 'unlucky' – when those topics that had been diligently revised come up, or indeed don't. Academics refer to this as the 'validity' of exams, and it is very well-studied. But of no account in England.

183. There is another matter too, a matter that is far less well-known, but which is based on something quite familiar.

184. In contrast to exams designed around a sequence of multiple-choice questions, in which a student simply ticks a particular box, exams in England require students to express themselves in their own words, so presenting their thinking: for example, by writing an essay on a particular topic, or by explaining scientific concepts, or as is evident in how a maths problem is tackled.

185. And we all know that, when a 'free-form' answer is marked, different, equally-qualified, examiners can give the same answer (slightly) different marks – for example, 11 or 12 marks out of 20. Neither of the examiners has made a mistake; neither has exercised their academic judgement inappropriately. The different marks are simply attributable to a legitimate difference in academic opinion.

186. The exam boards know this, and recognise it during their quality control of marking by associating each question with a pre-determined 'tolerance', this being the maximum number of marks by which any given mark is permitted to differ from a subject senior examiner's 'definitive' mark.

187. To allow an exam board to check on marking, one of the answers being marked might – unknown to the examiner doing the marking – have previously been given a 'definitive' mark, say 12 out of 20, by a subject senior examiner. If the 'tolerance' associated with that question is 2 marks, then a mark of 10, 11, 12, 13 or 14 is accepted as legitimate, but any other mark triggers an intervention by the exam board.

188. When all the marks for an exam are aggregated, it is therefore possible for the total to be, say, 54/100 or perhaps 56/100, depending on who did the marking. Importantly, both the marks 54 and 56 are equally legitimate, for each of the individual questions has been marked within 'tolerance'.

189. After all the marking has been completed, the exam boards, under the guidance of Ofqual, set the grade boundaries. Suppose in this example that GCSE grade 5 is defined as 'all marks from 53 to 57 inclusive'. In this case, the student receives a certificate showing grade 5 whether the mark is 54 or 56.

190. But if grade 4 is defined as 'all marks from 50 to 54 inclusive' and grade 5 as 'all marks from 55 to 59 inclusive', then the candidate's certificate shows grade 4 (corresponding to 54) or grade 5 (corresponding to 56), depending on the lottery of who did the marking.

191. That's problematic. And raises the question "which is the right grade?" – a question that Ofqual answers by reference to the 'definitive' grade, as determined by the 'definitive' mark given by a subject senior examiner. In this case, neither the examiner giving 54, grade 4, nor 56, grade 5, happen to be senior examiners, so the matter is unresolved. Suppose, though, that a subject senior examiner re-marks the script (for example, as a result of an appeal), and gives 55 marks. That confirms grade 5 as the 'definitive' grade, and grade 4 as 'non-definitive'. If the student awarded grade 4 appeals, a re-mark by a senior examiner therefore results in an up-grade to grade 5, implying that the originally-awarded grade 4 was 'wrong', and that the newly-awarded grade 5 is 'right'.

192. The fact that 'tolerance' is built-in to the quality control of marking, but ignored when grades are determined, invites the questions "in any year, how many legitimate, but 'non-definitive', grades, are actually awarded?", and "how effective is the appeals process in correcting the resulting errors?".

193. If the answers were "very few – the award of 'non-definitive' grades is very rare", and "very effective – the appeals process corrects very nearly all the errors", then that builds

confidence that exam grades are reliable and trustworthy, and that the exam system is (subject to the limitations mentioned in paragraphs 181 and 182) "fair". But if not…

194. These questions are not hypothetical, and have been researched by Ofqual. The answers are alarming. Firstly, the truth is that about one grade in every four, as awarded, is 'non-definitive', or 'wrong', amounting to some 1.5 million 'wrong' grades every year (Sherwood, 2019a). And, since 2016, the appeals process has been designed deliberately to prevent their discovery and correction (Sherwood, 2020f).

195. The key evidence that 1 grade in 4 is wrong is contained in an Ofqual report, *Marking Consistency Metrics – An update*, published in November 2018, documenting some of the results of an extensive research project, carried out in 2014 and 2015, in which entire cohorts of scripts in each of 14 subjects were, in essence, double-marked, once by an 'ordinary' examiner (as in 'usual' marking), and a second time by a subject senior examiner (as if on appeal), whose mark, and hence grade, was 'definitive' (Ofqual, 2018). Each script therefore had two marks and two grades, enabling Ofqual to answer a question such as "for GCSE Geography, how many scripts are awarded the same grade according to the marks given by both the 'ordinary' and the 'senior' examiner, and how many are given different grades?".

196. The results were presented in the report's Figure 12, reproduced here as Figure 2:

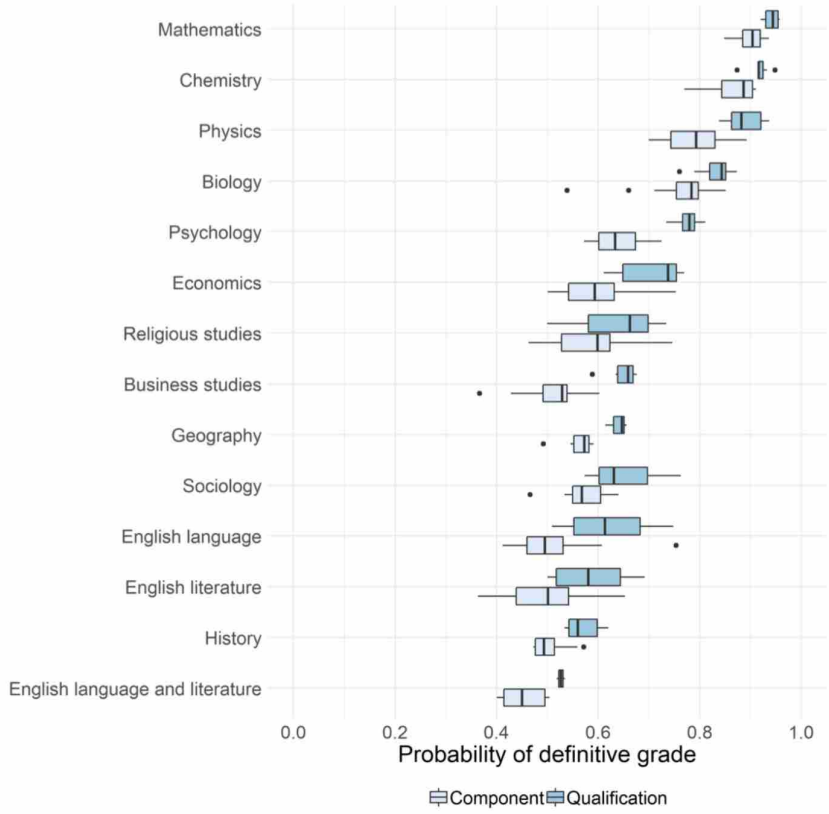


Figure 12. Boxplot showing the comparison of the probability of being awarded the 'definitive' grade at component and qualification level, for those GCSE, AS and A level qualifications for which we have full component data.[8]

Figure 2: Ofqual's evidence that 1 grade in 4 is wrong

Source: *Marking Consistency Metrics – An update*, Ofqual, (Ofqual, 2018)

197. For each of the 14 subjects shown, the important feature is the heavy black line towards the centre of the darker blue box. The position of this line, as measured along the horizontal axis, answers the question "For the aggregate of GCSE, AS and A-level exams in this subject, what fraction of the total cohort are awarded the 'definitive' grade?"

198. So, for example, for (all varieties) of Maths, that line correspond to 0.96. That implies that 96% of all Maths grades, as awarded each year, are 'definitive'/'right' and that the remaining 4% are 'non-definitive'/'wrong'. For Geography, the figures are about 65% 'right', 35% 'wrong'; for History, about 56% 'right', 44% 'wrong'.

199. Ofqual do not give an all-subject average. However, using numbers inferred from this chart, combined with knowledge of the corresponding subject cohorts, and making sensible assumptions about the subjects not reported (including, notably, all modern foreign languages, computer science, art, music and drama), a sensible estimate of that average is about 75% 'right', 25% 'wrong', or thereabouts (Sherwood, 2019a). That's the evidence that 1 grade in 4 is wrong.

200. As regards appeals, in May 2016, Ofqual changed the rules, explicitly denying any appeal on the grounds of legitimate differences in academic opinion – dismissed by Ofqual's claim that "*it is unfair to have two bites of the cherry*" (Ofqual, 2016). Even if the first bite is poisoned, and the second contains the antidote. For, as explained in paragraphs 183 to 199, the fact that 1 grade in every 4 is wrong is totally attributable to the consequences of legitimate differences in academic opinion.

201. That about 1 grade in 4 is wrong has been known by Ofqual, the DfE, and the exam boards since November 2015 (before the appeals rules were changed), as evidenced by an Ofqual Board Paper, dated 18 November 2015, confirming that the "*the report on potential quality of marking metrics is complete*" (Ofqual, 2015). And although Ofqual have done nothing to fix this problem, as well as being loathe to admit it, an important acknowledgement that this is indeed the case is the evidence given by Ofqual's then newly appointed Chief Regulator, Dame Glenys Stacey (Sally Collier's successor) at the post-Covid hearing of the Commons Education Committee on 2 September 2020. When asked by Ian Mearns MP about whether or not 25% of 'real' exams grades are unreliable, Dame Glenys did not refute the question, but replied that grades are "*reliable to one grade either way*" (Commons Education Select Committee, 2020c). This is not quite the same thing, but it is very close (Sherwood, 2020g).

202. The known 1-grade-in-4-is-wrong unreliability of 'real' exams is relevant to the Covid Inquiry in two key respects. Firstly, the mantra that "*exams are the fairest way*" is false, and has been known by the authorities to be false for years. The truth is that 1 exam grade in every 4 is wrong: wrong throughout the 2010s, and wrong now.

203. Accordingly, in 2020, *any process that might deliver results more reliable than 1 wrong grade in every 4 would have delivered more reliable results than any exam.* In this context, my vote goes for teacher judgement, especially teacher judgement carried out carefully, with integrity, and under appropriate supervision, all of which were quite possible. The 'bar' – 1 'real' exam grade in every 4 is wrong – is woefully low, so teachers

don't have to be particularly good to clear it. Indeed, as I will discuss in paragraphs 217 to 234, perhaps the TAGs actually did.

204. Secondly, I draw attention to some key evidence that emerged on A-level results day, 13 August 2020, when Ofqual (finally) published details on the algorithm in a highly technical, and dense, document running to 318 pages (Ofqual, 2020k). This chart, on page 81, is of especial importance:

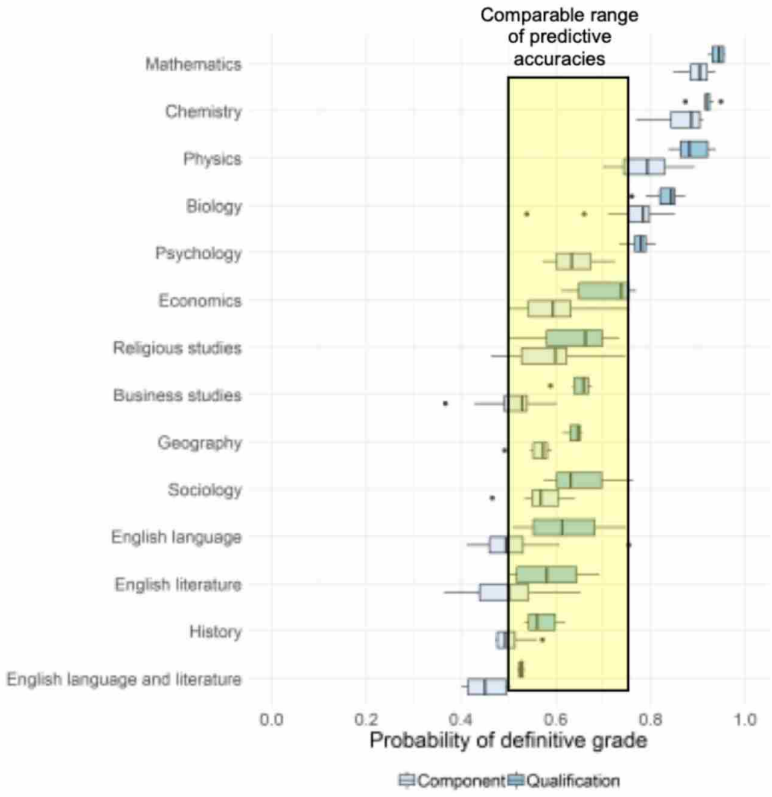


Figure 7.25 Probability of definitive grade being awarded based on an analysis of marking consistency reproduced from Ofqual 2018.

Figure 3: How the 2020 algorithm was tested

Source: *Interim report* (Ofqual, 2020k)

205. This chart was central to the method Ofqual used to test the algorithm's results, and to verify that the results of the algorithm could be trusted for every grade, as awarded to every student in the country. According to the algorithm's specification document, Ofqual's objective was to ensure that the algorithm delivered grades that fell within the shaded region to the right of centre of this chart, designated the "comparable range of

predictive accuracies". If the algorithm could achieve that, then the algorithm was good to go.

206. As is evident, however, the 'target' the algorithm was being designed to achieve is based on the chart shown as Figure 12 in Ofqual's 2018 report, as reproduced as Figure 2 in this submission, and as discussed in paragraphs 195 to 199.

207. That chart, however, is Ofqual's own evidence that 1 exam grade in every 4 is wrong.

208. That implies that the intended objective of the algorithm was to predict grades with a reliability of 75% at best!

209. I find that breathtaking.

Oh, what a missed opportunity!

210. In March 2020, when exams were cancelled, Ofqual and the other authorities knew that one 'real' exam grade in every four, as awarded on candidates' certificates, was wrong. And had been wrong for years. But clearly they didn't mind, for not only had they done nothing about it, they'd deliberately covered it up when, in 2016, they changed the rules for appeals to prevent their discovery and correction.

211. In the absence of exams, Ofqual chose to develop and use an algorithm to determine how many students should be awarded each grade in each subject at each school. As an objective, that, to me, is combination of supreme arrogance and supreme ignorance in that no algorithm could ever do that. But my opinion is coloured by my assumption that the algorithm was intended to give the 'right answers'. According to the chart reproduced in paragraph 204, however, it appears that Ofqual had no intention to get the grades 'right' – as long as three grades out of four were right, to Ofqual, that was good enough. Oh dear.

212. Yet Gavin Williamson, in the DfE press release of 20 March, said "*I have asked the exam boards to work closely with teachers who know their students best to ensure their hard work and dedication is rewarded and fairly recognised*" (Department for Education 2020a). Yes, those are the right words. But the reality was so, so different.

213. That reality, though, didn't have to be different. Gavin Williamson's words could have become real, for it is not difficult to envisage – and then design – a process whereby

teachers would indeed use their experience, their knowledge, and their professional judgement to determine their students' grades. Certainly, this opens the doors to bias (both ways), malpractice and laziness. But those doors could easily have been policed – for example, by 'external examiners', whereby schools collaborated to 'share' their teachers; by having regional 'experts', nominated by the teacher unions (ASCL, HMC, NAHT, NEU…) to oversee matters in their locality, by being alert to – but not constrained by – historical patterns to help avoid giving all the students an A*. Furthermore, there was the possibility of trialling different approaches in different parts of the country, and learning accordingly.

214. Had that happened, the experience would have had enormous value, potentially proving that teacher assessment can work, and work well, giving results that are more reliable than 1-grade-in-4-is-wrong 'real' exams. Had that happened, that might have transformed the way students would be assessed in the future, making the entire assessment process – in my opinion – more valid, richer and more reliable.

215. Alas, that isn't happening, for – as noted in paragraph 16 – the Interim Report of the Curriculum and Assessment Review, chaired by UCL Professor Becky Francis, published in March 2025, states that "…*externally set and marked exams are an important way to ensure fairness as part of our national qualification system*…" (Curriculum and Assessment Review, 2025). To me, that's a variant of "*exams are the fairest way*", so it looks as if exams, as they are currently taking place, will continue, as they are now, into the indefinite future. Which includes the fact that one grade in four is wrong, for the reliability and trustworthiness of grades is not mentioned at all in the Interim Report.

216. So the events of 2020, and 2021 too, which provided a second opportunity to get things right, are to me, a great lost opportunity.

Were the TAGs more reliable than exam grades?

217. Despite the lack of a coordinated, well-planned, and effectively supervised way of doing things, did the process of 2021 'get things right' – or, more pragmatically, deliver grades more reliable than 1-grade-in-4-is-wrong 'real' exam grades?

218. To my knowledge, this has not been studied. But might I offer a hint that perhaps, in 2021, the TAGs were more reliable than exam grades, a hint based on my own research,

conducted over the last several years, on the reliability – or rather unreliability – of 'real' exam grades.

219. To help with this research, I have developed some spreadsheets that enable any 'real' grade distribution to be analysed in terms of three important features:

- The mean, or centre, of the distribution of marks, using a scale from 0 marks to 100, and assuming that the distribution is what mathematicians call 'Gaussian', or 'normal', as recognised by its characteristic 'bell' shape.
- The standard deviation of this distribution, so determining the width of the 'bell'.
- The positioning of all the grade boundaries, for example, such that grade 4 is 'all marks from 50 marks to 54 marks inclusive' and grade 5 'all marks from 55 to 59 inclusive'.

220. Using these spreadsheets, I am able to simulate, in detail, the grade distribution for any historic GCSE or A-level examination, and, in so doing, to estimate how many students were awarded each of the marks, 'normalised' on a scale from 0 to 100. This is important, for it is an estimate of the distribution of 'raw' marks as given by the examiners, this being the distribution over which grade boundaries are placed, and from which the grades, as shown on candidates' certificates, are subsequently determined. This distribution of 'raw' marks then provides a basis for exploring, for example, how the grade distribution might be different if the grade boundaries were set in different places, or what might happen if Ofqual were to use a different policy for determining a student's grade from the underlying mark. Let me emphasise my willingness to make my spreadsheets available for scrutiny.

221. As an example, Figure 4 shows the results of using my spreadsheet to simulate the overall results for the total of 5,070,841 grades actually awarded for the 2019 GCSE exams in England. The actual percentages of that total awarded each grade in 2019 are depicted as the grey columns (on the left of each pair), and the corresponding percentages as calculated by my spreadsheets are shown as the hatched columns (to the right):

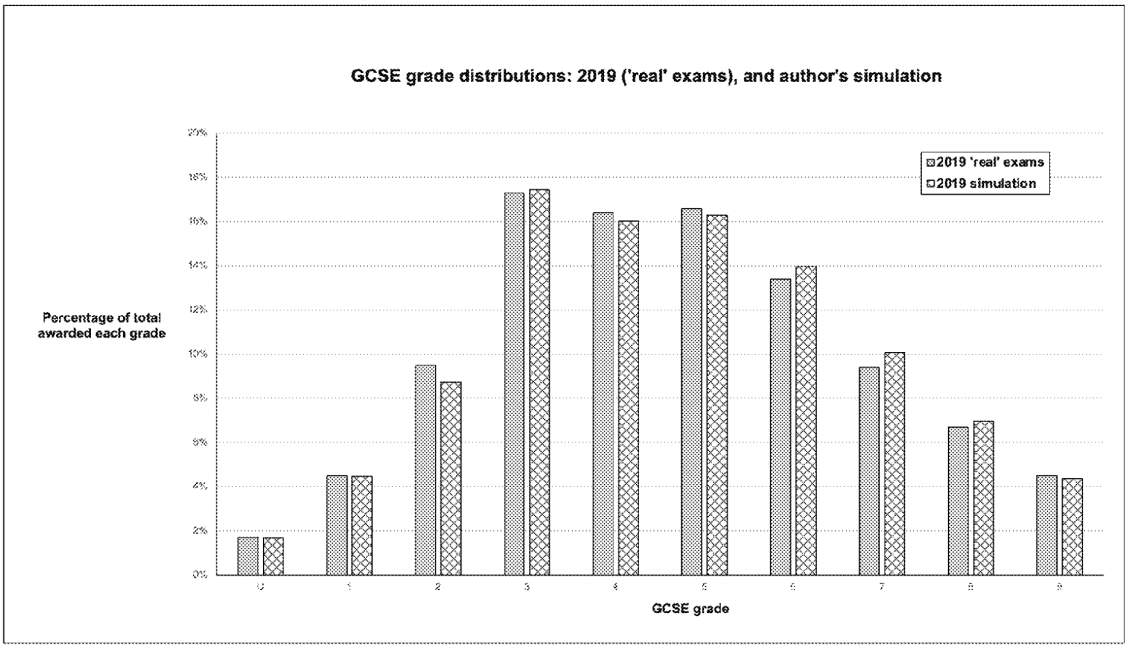


Figure 4: My simulation of the actual 2019 'real' grade distribution

222. As can be seen, the 'fit' is not perfect: my calculation over-estimates grades 6, 7 and 8 somewhat, and under-estimates grade 2. That said, the overall shapes are quite similar.

223. Importantly, the spreadsheets 'discover' the locations of the grade boundaries that resulted in the 2019 grade distribution, so enabling me to apply those grade boundaries to the 2021 GCSE cohort. This therefore answers the question "For the 5,236,861 GCSE grades awarded in 2021 (Joint Council for Qualifications (JCQ), 2024), how many correspond to each mark on a scale of 0 to 100, resulting in the same grade distribution as the 2019 'real' exams?". That therefore gives insight into what would have happened if the 2021 cohort had taken 'real' exams, marked and graded to the same standards, and according to the same policies, as were used in 2019.

224. One of these policies is 'no grade inflation', so constraining the overall grade distribution to be similar to the distributions of previous years; another is the policy for determining grades from the 'raw' marks, such as 'grade 4 is defined as corresponding to all marks from 50 to 54 (out of 100) inclusive' and 'grade 5 is all marks from 55 to 59 inclusive'. That may seem to be an 'obvious' policy, as indeed it is. But, as I now discuss, it is not

the only possible policy for determining grades from 'raw' marks, especially in the light of the particular weakness that this policy results in unreliable grades.

225. For as discussed in paragraphs 186 to 191, it is possible that a script legitimately marked 54, and so awarded grade 4 under this policy, might have been given the equally legitimate mark 56 had the answers been marked by a different, equally-qualified, examiner. A mark of 56, though, corresponds to grade 5 – hence the unreliability of grades as examined in paragraphs 183 to 199, in which it was explained that the unreliability is attributable to the failure of Ofqual to recognise the concept of 'tolerance' in the awarding of grades.

226. A feature of my simulation is the opportunity to explore different policies for grading, such as policies that incorporate 'tolerance'. One such policy is this: instead of determining the grade of a script marked 54 from the mark 54, to determine the grade from the mark 54 + 2 = 56, where the addition of the extra two marks recognises the possibility that another examiner might, legitimately, have given that script 56 marks. Although, at first sight, this appears to be startling, it has a highly beneficial consequence. As explained in detail elsewhere (Sherwood, 2022, pages 338 - 351), by recognising the 'tolerance' present in the original marking, and by giving the candidate 'the benefit of the doubt', the likelihood that a re-mark by a senior examiner would result in a grade change is much reduced. This therefore improves the reliability of the originally-awarded grade, so increasing the grade's trustworthiness.

227. Adding two marks is not the only possibility – adding one, if the 'tolerance' for a particular subject is narrower, or adding three, if the 'tolerance' is broader, are alternatives. And there are many other, different, grading policies too (Sherwood, 2019b, and Sherwood, 2022, pages 324 - 363).

228. If this 'benefit of the doubt' concept is put into practice on the 2021 TAG data, using the same grade boundaries – and hence applying the same academic standards – as in 2019, the results are as presented in Figure 5, in which:

- The TAG distribution, as actually awarded in summer 2021, is shown by the central black columns.
- The results of my simulation, applying the same standards as 2019 but grading according to a policy of adding 2 marks ('$m + 2$'), are shown by the light grey columns to the left of each triplet.

- The results of my simulation, applying the same standards as 2019 but grading according to a policy of adding 3 marks ('$m + 3$'), are shown by the dark grey columns to the right of each triplet.

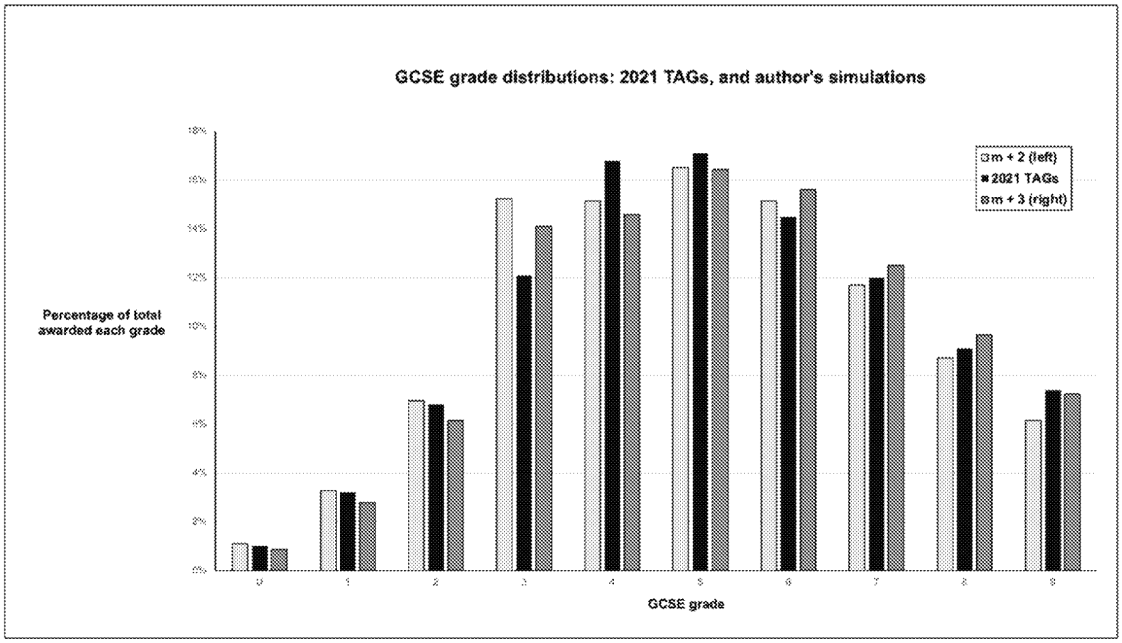


Figure 5: Were the 2021 TAGs more reliable than 'real' exams?

229. As can be seen, the '$m + 2$' and '$m + 3$' simulations straddle the TAGs, with the '$m + 2$' simulation somewhat overestimating grades U, 1, 2 and 3 and underestimating grades 5, 6, 7, 8 and 9, whereas the '$m + 3$' simulation does largely the opposite, underestimating grades U, 1, 2, 4, 5 and (marginally) 9, whilst overestimating grades 3, 6, 7 and 8. That both simulations overestimate grade 3 and underestimate grade 4, and to a lesser extent grade 5, suggests to me that, in determining the TAGs, teachers were in general 'generous' at the 3/4 boundary, awarding fewer grades 3 and more grades 4, so ensuring that fewer students 'fail', especially in those critical (and large cohort) subjects of English and Maths (see paragraph 168). That all said, the 'fit' between my simulations and the actual TAG results is quite good, and would probably be even better for an intermediate adjustment such as '$m + 2½$', if that were possible - for, like students, marks usually come only in whole numbers.

230. The basis of both simulations, shown as the light and dark grey columns in Figure 5, was a replication of applying the 'real' exam-based, 'no grade inflation', grade

distribution of 2019 to the 2021 cohort, resulting in a simulated distribution very similar (as shown in Figure 5) to the actual distribution of 2019's actual exams – exams that were 75% reliable, in that 75% of the grades awarded in 2019 were 'definitive'/'right' and 25% were 'non-definitive'/'wrong' (see paragraph 199).

231. Then, keeping the standards for marking as used in 2019, but simply changing the policy for grading from determining the grade, as in 2019, directly from each script's mark to a determination based on that mark plus 2, the result is a distribution as shown by the left-hand light grey columns - a distribution associated with grades that are about 90% reliable. By determining the grade by adding 3 marks, the result, shown by the right-hand dark grey columns, is associated with a reliability of about 96%.

232. Given that these two simulated distributions closely match, and straddle, the TAG distribution, with the '$m + 2$' distribution being somewhat closer, might this suggest that TAGs were perhaps 91%, or maybe 92%, reliable, as compared to the 75% reliability of 'real' exams? Did teachers actually do a better job in 2021 than exams have ever done, let alone the 2020 algorithm?

233. Figure 5 shows the results of my study for the GCSE TAGs; I have obtained similar results for the A-level TAGs, and for both the GCSE and A-level CAGs – let me reiterate my willingness to make all details available for scrutiny if that might be helpful.

234. So, with this evidence in mind, maybe teachers really can be trusted…

Statement of Truth

I believe that the facts stated in this witness statement are true. I understand that proceedings may be brought against anyone who makes, or causes to be made, a false statement in a document verified by a statement of truth without an honest belief of its truth.

Signed: Personal Data

Dated: 18 August 2025