Written evidence submitted by Dennis Sherwood

Written evidence from Dennis Sherwood in connection with understanding why so many CAGs were over-bid, and to suggest an approach to the award of fair grades

I submit this on the evening of 12th August, before the announcement of the A-level results tomorrow, and some 24 hours after the announcement by Gavin Williamson that an appeal may now be made against the awarded grade on the grounds that the corresponding mock grade is higher.

The context is the knowledge we now have that many CAGs have been over-bid. Is this because teachers have been 'over-optimistic'?

That is a possibility. But not the only one. There could be a much more prosaic reason – a reason that can be traced back to Ofqual's failure to design a wise process, and to give schools clear and complete instructions as to how to operate it.

My submission comprises a [blog published on the HEPI website](#) earlier today, 12th August 2020.

The Great CAG Car Crash – What Went Wrong?

Dennis Sherwood

As a result of the [public uproar](#) following the ['adjusting down' of around 124,000 centre assessment grades (CAGs)](#) – about [one-quarter](#) of all grades submitted – Scotland's Education Secretary, John Swinney, has now [binned "statistical standardisation" and reinstated schools' down-graded CAGs](#). In England, the numbers aren't known yet, but a [recent report](#) produced compelling evidence, based on (the [somewhat suspect](#)) Slide 12 from Ofqual's recent [Summer Symposium](#), that about 40% of A-level CAGs will be down-graded. This too is driving a [build-up of public pressure](#), the final outcome of which is as yet unknown.

John Swinney also announced the inevitable [enquiry](#) into what went wrong, including an autopsy of the process, as well as trying to get to the bottom of why so many CAGs were over-bid, for which two explanations are already on the table: ["over-optimistic" teachers](#) and [discrimination against the socially disadvantaged](#).

But are these the full story?

The muddle of the over-bid CAGs needs to be untangled not only in Scotland, but in Northern Ireland, Wales and England too. And to do that, may I suggest that someone should look in detail at the relevant evidence, the CAGs, and ask what to me are these two key questions:

- "How many of the CAGs were submitted in good faith and were plausible?"
- "How many appear to have been submitted by fraudsters, chancers, game-players, and the lazy?"

I'll deal with the second question first. If the CAGs submitted by any school are way higher for the top grades than the school's subject history, that's evidence of, let's say, game-playing. So, for example, a teacher who thinks "I can't be bothered with all this. I'll just submit top grades and let the board sort it out". Or someone who, fearing confrontation with irate parents, decides to submit A*s and 9s for everyone – that way, the teacher can look any parent in the eye and say, "I submitted a top grade! It's not my fault the outcome was [whatever]! Blame the exam board, not me!".

Such heavily distorted submissions should be easy to spot, and I trust that there will be very, very few.

The first question, about plausible submissions, requires rather more explanation. Anyone who tried to produce this year's CAGs will have hit two, apparently trivial, but in fact potentially devastating, arithmetical problems: rounding and historical variability.

Suppose, for example, that the appropriate historical average is that 30% of previous students were awarded, say, grade B. This year's cohort is 21 students. 30% of 21 is 6.3. That's not a whole number, which is a problem: students don't come as decimals, but as whole numbers. So the teacher faces the dilemma of rounding down to 6 or up to 7. The rules of arithmetic say 'round down'. But that 7th ranked student is quite good, and really deserves a B, so let's submit 7. So reasonable; so human; so understandable.

But if, in good faith, teachers in many schools rounded up, then grade inflation is blown sky high, for this is the 'Tragedy of the Commons'. To maintain "no grade inflation", there must be as many roundings down as up, which is most unlikely.

There's another consequence of rounding too, best illustrated by a rather odd-looking example, but it does make my point.

A school's historical average grade distribution is such that 10% of its students were awarded each of the ten grades 9 – 1 and U. This year's cohort is 9. So that's 0.9 of a student in each of the 10 grades, each

rounded to 1. When I add the rounded figures, the total cohort is 10. But there are only 9 students. Where did that extra 'student' come from? From the accumulated rounding errors. And so to correct for that, I have to deduct one 'student'. But which one? From which grade? From grade U of course. That way, each of the 9 students in the cohort are awarded each of grades 9 – 1, with no award of the U.

That makes sense. But there was a choice: I could have awarded one each of grades 8 – 1 and the U. But why on earth would I? And if everyone in a similar position chooses the highest grade, not the lowest, guess what happens to grade inflation...

One more example.

Suppose that, historically, the percentages for grade B were 40%, 20% and 30% over each of three previous years, which – since the cohorts are the same size in each year – average to the 30% used earlier.

If, instead of using the average, I use the best of these years – after all, this year's cohort is just as good as that one, if not better – then 40% of 21 is 8.4, which I'll round up to 9. That's good – I'll submit 9, that's sure to be fine.

But alas no.

Submitting the rounded up 7, or the 9, or a compromise of 8, could all create havoc if everyone does the same. And why shouldn't they? It's all very reasonable...

..especially since neither the SQA nor Ofqual specified the rules!!!

If teachers had been instructed how to do the rounding; if teachers had been instructed just how close they had to be to the average; if teachers had been given the same calculation tool that looked after all this techy stuff consistently and 'behind the scenes'; then they might have submitted CAGs-that-the-algorithm-first-thought-of, these being the 'right answers'. And even better if they had also been allowed to submit well-evidenced outliers.

But in the absence of these rules, teachers were aiming at moving goalposts in the dark. No wonder there have been so many misses.

My thesis is that 'plausible overbids' are not the fault of the teachers. To me, the blame lies totally at the door of the SQA and Ofqual for not making the rules clear. Chancers and fraudsters are another matter, of course.

I think that 'plausible' and 'gamed' over-bids can be untangled by seeking the evidence – by looking through the CAGs and discovering the patterns, as illustrated in the Figure. And I think this should be done with urgency.
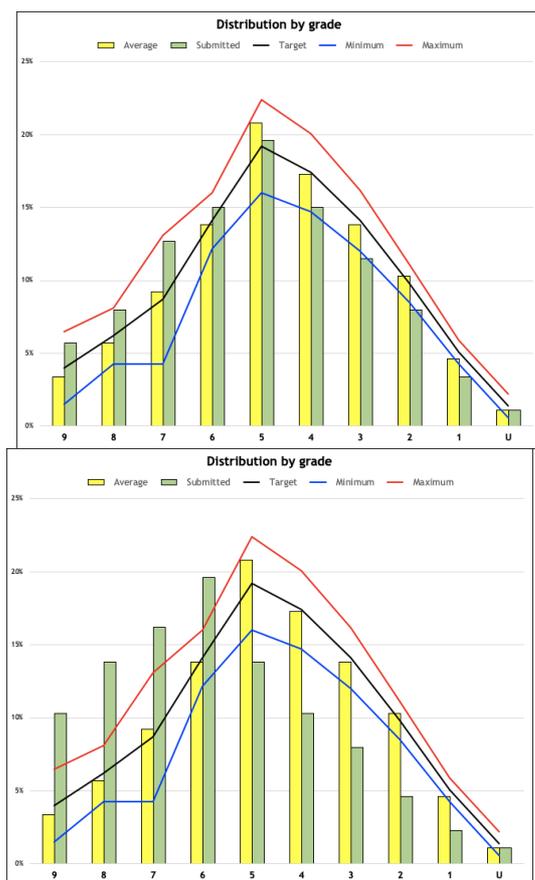
'Plausible' and 'gamed' grade distributions

In these hypothetical examples of the distribution of GCSE grades for the same subject cohort, the central black line is the historic average; the upper red line, the historic maximum; the lower blue line, the historic minimum. For the 2020 cohort, the distribution that most closely fits the historic average is shown by the yellow columns.
The green columns show the grades as submitted.

On the left, no submitted grade exceeds the maximum, and only grade 1 is just below the minimum. Such a distribution is, in the context of the article, 'plausible'. Pause for a moment to guess the grade inflation implied by submitting the 'plausible' green distribution rather than the exact average yellow distribution. The answer is nearly 6 percentage points. The % 9 – 4 of the yellow distribution is 70.2%; of the green distribution, 76.0%.

On the right, the higher grades all exceed the maximum; the lower grades are all below the minimum. This is the typical pattern of a 'gamed' distribution.

I have no idea what the outcome might be. Perhaps most of the over-bids will be shown to be attributable to fraudsters and gamesters; perhaps not.

In Scotland, the decision has been taken to scrap the algorithm's results, and to accept schools' CAGs, even if they really were over-the-top (but, hopefully, in only a few cases…).

In England, the grades to be announced shortly will, subject to the "small cohort" rule, be those determined by the algorithm, as they have always been. What has been changed by Gavin Williamson's last minute announcement is a tweak to the rules for appeals.

Until last Thursday (6th August), the grounds for appeal were limited to technical and procedural errors. On that day, and after much pressure, the rules were widened to allow appeals if schools "can evidence grades are lower than expected because previous cohorts are not sufficiently representative of this year's students".

Last night (11th August) came the news that the grounds for appeal had been amended a little more: schools can now appeal their awarded grades if their students' mock results are higher. I'm puzzled by that. If an alternative to calculated grades is to be used as a criterion of "right/wrong", why choose mocks when the CAGs are immediately and easily available, and already have mock results factored in? And not just mock results: Pages 5, 6 and 7 of Ofqual's Guidance notes, for example, list all the aspects of student performance that CAGs were to take into account. Are all these of no value? Has all this important evidence been discarded? Have mocks been chosen in preference to CAGs because the CAGs are all wildly 'over-optimistic' and just can't be trusted?

But as I hope I have demonstrated, some CAGs might not be 'over-optimistic', but rather 'plausible'. We just don't know. And I think we should find out.

For if we did, that might provide another way out of this appalling mess.

Suppose, for a moment, that all the English CAGs are reviewed to determine which are 'plausible' and which are 'gamed'. Suppose further that Ofqual adopt the rule that all CAGs that are 'plausible' are either confirmed (if already awarded) or re-instated (if they have been over-ruled by the model). To complete the picture, those CAGs that have been 'gamed' would be over-ruled by the model (as may well have already happened). And since some students of 'gaming' teachers might have been penalised by the award of a calculated grade, there also needs to be a free appeals process, open to any student who feels he or she has been awarded an unfair grade, and who can provide suitably robust evidence, of which mock results can be one element.

This will certainly drive some grade inflation – but I would argue that this is a consequence of Ofqual's abject failure to design a wise process. The guardian of the "no grade inflation" policy is totally responsible for its breach.

September 2020