Written evidence submitted by Professor John Gray

A MUCH SIMPLER WAY OF AWARDING EXAM RESULTS IN EXCEPTIONAL CIRCUMSTANCES

Professor John Gray
University of Cambridge

A Response to the Select Committee's Call for Evidence on: The effects of cancelling formal exams, including the fairness of qualification awards and pupils' progress to the next stage of education or employment.

John Gray is Emeritus Professor of Education at Cambridge University and former Vice-Principal of Homerton College. A Fellow of the British Academy, he chaired the Standards Committee at Cambridge Assessment for more than a decade and was a Special Adviser to the House of Commons Select Committee's investigation into School Accountability in 2009-10.

Summary: The paper presents a straightforward approach to the setting of exam grades which could have been used during the summer 2020 session. Based on centre-moderated grades, and historic patterns of achievement, it has two main advantages. First, it is likely to be perceived as considerably fairer by teachers and pupils. And second, it meets concerns about so-called 'grade inflation'.

Lessons from the past

Three major difficulties seem to have been encountered during the summer 2020 session:

1) Confusion and misunderstanding about the role teachers' predictions and 'centre assessed grades' would play in the final determination of the results for individual pupils.
2) Overly-optimistic assumptions about the ability of a statistical model (the so-called 'algorithm') to provide credible results for individual pupils.
3) The maintenance of organisational mind-sets about how to proceed which were largely premised on resolving potential anomalies rather late in the day.

Teachers as Examiners

Chapter 3 of Ofqual's report (2020) provides a good account of the strengths and limitations of teachers' contributions (both actual and potential) to the business of marking and grade setting. Reviewing the evidence and commenting on how 'accurate' teachers' estimates can be (Ofqual, 2020, para. 3.1), it distinguishes two dimensions:

- absolute accuracy: the ability of a teacher to estimate the actual grades that individual students will achieve; and

- relative accuracy: the ability of a teacher to estimate the rank order of students by their grades, ie their levels of achievement relative to each other

Teachers turn out to be rather good at making judgements which correspond with the examiners' rank orders of pupils' performance – the correlations between their judgements and those of the examiners are impressively high (Ofqual, 2020).

By comparison, they are less good at predicting the *actual* grades pupils will be awarded – the correlations are a good deal lower (Ofqual, 2020, pp.13-16). This is perhaps less surprising as the process by which examiners determine the cut-offs between grade boundaries for different grades only emerges rather late in the day; indeed, prior to final standardisation meetings most examiners themselves may be a bit uncertain how exactly marks will be turned into grades.

There is an additional ambiguity about the ways in which teachers' predicted grades are developed. Predicted grades serve multiple purposes. They are used, amongst other things, to provide information on which HE providers and future employers may make judgements about students' potential as well as to motivate pupils by 'stretching' them. In both cases an element of optimism is built into the process. In the circumstances it is not particularly surprising that teachers' predictions, generated under these conditions, do not correlate as strongly with the grades examiners eventually award.

On the other hand there is growing evidence that teachers have become increasingly sophisticated in judging their students' performance, not least because they have had an array of information about students' prior attainment available to them (Ofqual, 2020, pp. 15). To the best of my knowledge, however, no research study has asked teachers to predict what grades their students will actually be given as opposed to those they feel they deserve. In an ideal world the two may be identical.

In short, in exceptional circumstances such as those faced in Summer 2020, the most valid form of assessment is likely to be one which gives teachers a major role in determining the eventual outcomes but subjects their judgements to various challenges and constraints.

Benchmarking the Standard

There has rightly been concern about 'grade inflation' diminishing the credibility of exam qualifications. In other words, the larger the number of top grades, the more sceptical the wider audience may become about their value. The challenge therefore is to develop benchmarks which are fair to a particular cohort of pupils but do not stretch the bounds of credibility. As Ofqual has identified, the school's own track record in recent years is probably the most obvious reference point.

There are problems however with this approach. Understandably, many schools will want to claim that they are on rising improvement trajectories and that their results are constantly getting better. Ofqual investigated schools' trajectories and concluded that in the greater majority of cases results tended to be stable from one year to the next; only a very small minority (less than one per cent) seemed to be on modestly rising trajectories when compared with previous years' results (Ofqual, 2020, p.21).

These conclusions are supported by earlier independent research (Gray et al, 2001;Thomas et el, 2007; Mangan et al, 2007). Within fairly narrow bounds, schools' performances tend to bounce around a bit from year to year. In the majority of cases it is almost impossible to extrapolate next year's results from knowledge of this year's within any degree of accuracy. They may go up a bit, they may go down a bit. Some schools, however, do seem to manage to improve their performances year on year, taking into account the changing nature of their intakes. To talk of a trend, however, requires a minimum of three years of upward movement. Only a very small minority of schools manage to achieve this. Most importantly, perhaps, it is fairly rare for a school to sustain improvement into a fourth or subsequent year.

In short, these patterns suggest that it is not unreasonable to use the last three years of a school's performance as a reference point. This is in fact what the Ofqual model did.

Knowledge of year on year variations is helpful but it may not count for much in the eyes of a pupil who believes that they have been disadvantaged by a 'bad year' amongst previous cohorts in their school. To deal with this concern, in exceptional circumstances, it is important to be seen to be generous in determining the appropriate benchmarks. Allowing schools themselves to choose their benchmark therefore commends itself.

Schools will doubtless choose their 'best year' in the last three to benchmark their performance. In many cases however choices about which is 'the best' will not be self-evident. They can be reassured that the worst that will happen to them is that they plateau or mark time. To smooth out fluctuations in small subjects a decision rule would probably be needed that required the results of any two out of three years to be combined. The fact that that each school will have made its own choice will considerably increase the likelihood of the benchmark being perceived as legitimate.

The Various Steps Towards Determining Final Results

1) ***Establishing initial benchmarks against which to reference performance:*** Exam boards would ask schools to nominate their benchmarks and to submit them for verification. By virtue of schools nominating their own benchmarks there is likely to be small increase in results because everyone will be choosing their 'best' year. However, given what we know about year to year variations, this is likely to be comparatively small. To minimise the burdens, and unlike in normal years, schools would not be asked for teachers' *predicted* grades; it might also be possible to forgo asking them to provide rank orders of pupils subject by subject.

2) ***The national picture:*** The exam boards would construct the overall national picture from the schools' submissions. Early on in the process, then, Ministers could be informed about likely outcomes and could determine whether they are likely to be satisfied by the total picture or whether they wish impose any constraints in the interests of year-on-year fairness and comparability across cohorts. There might, for example, be some evidence that the cohort in question had markedly lower prior attainments than in previous years for which some sort of adjustment might be

thought appropriate; however, minor differences are probably best left unaccounted for.

3) ***Profile for each individual school:*** Based on the schools' benchmarks, the boards confirm to schools the expected distribution of grades, factoring in the size of this year's cohort. Schools will already have a reasonable idea of what to expect from knowing what they submitted for validation.

4) ***Centre-moderated assessments:*** Each participating centre is asked to set up an Assessment Committee to oversee the setting of grades. Individual subject departments would be informed what the expected distribution of grades would be and asked to make initial recommendations about the grades to be awarded to individual pupils. These judgements would then be debated and signed off by the Assessment Committee. This is where the teachers' knowledge of individual learners would be significantly important.

5) ***Decisions at the grade boundaries:*** Assessment Committees will be faced with some dilemmas. Just as Board examiners have to make difficult decisions about grade boundaries, so too teachers may find that they have pupils on the borderline of particular grade categories they would wish to promote. The greater majority of schools will have further evidence they can factor into the discussions about each pupil at this stage. In the interests of facilitating fairness and legitima*cy*, it is suggested that Assessment Committees be allocated a small number of 'wild cards' (the exact number to be dependent on the number of entrants) which can be played to increase the number of grades available at a particular level in a particular subject above that proposed by their exam board. These 'wild cards' are likely to be mostly used at the borderlines of the higher grades and short reasoned justifications could reasonably be expected as part of the audit trail.

6) ***Confirmation of grades:*** Assessment Committees would be asked to provide a reasoned account of the processes they had used, the evidence available to them and any dilemmas they had experienced. At the same time they would submit their proposed grades to the exam boards for signing off. The boards would then issue their final awards.

7) ***Appeal processes****:* Pupils who were dissatisfied with their grades would have two (limited) avenues to pursue their concerns. First, they could draw their school's attention, for example, to things like the grades their teachers had predicted for them in HE or their post-school employment applications as well as other sources of feedback and marking. Second, they could be given the option of sitting a formal exam later near the start of the next term. Cases which alleged maladministration or other misdemeanours would continue to be submitted to the boards in the first instance.

References

Gray, J., Goldstein, H. and Thomas, S. (2001) Predicting the Future: the role of past performance in determining trends in institutional effectiveness at A level, *British Educational Research Journal*, 27, 4, 391-405.

Ofqual (2020) *Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report*, London: Ofqual.

Mangan, J., Pugh, G. and Gray, J. (2007) Changes in examination performance in English secondary schools over the course of a decade: searching for patterns and trends over time, *School Effectiveness and School Improvement*, 16, 1, 29-50.

Thomas, S., Peng, W.J. and Gray, J. (2007) Modelling patterns of improvement over time: value added trends in English secondary school performance across ten cohorts, *Oxford Review of Education*, 33, 3, 261-295.

September 2020