Written evidence submitted by Professor Guy Nason, Imperial College, London

Concerns over the process for awarding grades for summer exams in the UK.

I am writing to submit evidence to your inquiry into 'The impact of COVID-19 on education and childrens' services'.

I have concerns over the process, including statistical standardisation, that will be used to produce exam grades for Summer 2020. Although we do not yet know the details of the standardisation process, my concerns are based on statements contained within various Ofqual documents and the nature of the data that has already collected from exam centres.

First, though, the COVID situation that the UK finds itself in is, evidently, extremely difficult. At the start of the pandemic, we had little evidence about its nature and seriousness. With summer exams looming the Government bravely made the tough decision to cancel exams and move to a system of teacher-predicted grades. Ofqual has been given the difficult task of making this work for students, parents, teachers and the whole of the UK that relies on exam grades.

From the start, I have found it difficult to see how any process could be developed that would ensure the integrity of the system and be fair to everybody. I do think Ofqual has done a good job, based on the cards it was dealt. However, I think that there are issues that will arise, and these will undoubtedly lead to appeals within the system and, probably, legal action outside of it. Although I am, of course, keen on maintaining standards, I worry less about maintenance of the integrity of the system, but worry more about fairness to individual students. Hence, I would urge the Government, Ofqual and your committee to press hard on ensuring fairness to individual students.

My main concerns fall into two categories, which I will attempt to explain more fully below. They are that

1. there is a belief that teacher-predicted rankings are more accurate than they really are. In particular, Ofqual consultation documentation appears to selectively quote correlations between predicted and actual grades that are higher than are found, even in the literature that they reference.

2. ranking uncertainty is considerably underappreciated. This could lead to unfair decisions if, for example, exam boards move students between grades using ranking information, without taking the ranking uncertainty into account.

1. Reliability of Predicted Grades?

- The consultation document *Exceptional arrangements for exam grading and assessment in 2020*[1] asked a number of questions. A question on page 29 asked whether respondents agreed or disagreed with a particular approach to statistical standardisation. The question was preceded by descriptions of such an approach and included the line:

> *"research suggests that teachers can accurately rank order their students (correlations between the rank order of teacher predictions/ estimates and actual grades are relatively strong at around 0.76 to 0.85)"*

[1] Exceptional arrangements for exam grading and assessment in 2020. 15th April 2020. Ofqual/20/6610.

Unfortunately, that *research* was not referenced in the document. I asked Ofqual (by email on June 16th) what was the source reference that underpinned this, and they referred me (June 19th) to 'the most relevant reference' as Dhillon (2005)[2].

Dhillon's research indeed showed correlations ranging from 0.77 (for A level Psychology) to 0.85 (for both A level Chemistry and Maths) for six subjects, see her Table 1. However, her table also presented correlations for 26 exams from six earlier studies. The correlations for these ranged from 0.45 to 0.82. The lower quartile, median and upper quartile for all 32 correlations reported in her Table 1 were 0.61, 0.67 and 0.77, respectively.

Although there is a suggestion that correlations have increased over time, it appears to me that the range quoted in the Ofqual consultation question (0.76 to 0.85) is selective and possibly deceptive. Even the second most recent study, Baird (1997), reports a correlation of 0.63 and the one before that, Delap (1992), reports one of 0.59. It is not clear to me that teachers, students and parents would be happy relying on predicted grades with correlations of around 0.5–0.6.

- I also have statistical concerns about the use of correlations on categorical data (such as A, B, C, D, E grades) as quoted by Ofqual and carried out by Dhillon (2005). Dhillon (2005), page 71, herself acknowledges this:

  > *"While use of the test may be viewed as problematical in that the data do not represent the true interval measure appropriate to it, its frequent use in previous studies mitigates its use here for comparative purposes"*.

  My translation of this would be "we probably should not be using correlations on grades, but it's ok since everybody else has done it previously". More details on my concerns appear in the annex below (point 1).

- Unfortunately, even recent educational research literature uses correlations in this context, sometimes without acknowledging the statistical problems mentioned above. Further, there is a separate problem, relating to individuals. In these situations, correlation is a statistic calculated on a large number of individuals. Even a reasonably large correlation (e.g. 0.7) does not automatically mean that prediction for all individuals will be successful. We would like students to be properly assessed, but when this is not possible, we do not want predictions for <u>some</u> of them to be terrible due to chance. Teacher-predicted grades can be used to maintain the integrity of the overall system but can be disastrous for some individuals.

- Dhillon (2005) noted that exam centre type has a strong influence on the accuracy of teacher-predicted grades. She noted that *'selective and independent schools'* tended to predict most accurately and *'further education establishments'* the least. For example, she notes *'the Byzantine task FE teachers face when judging the effort and ability of new, unfamiliar and frequently older students, not uncommon within such establishments.'*, page 79. She also found that 'age group' 'significantly

[2] Dhillon, D. (2005) Teachers' estimates of candidates' grades: Curriculum 2000 Advanced Level Qualifications. *British Educational Research Journal*, 31, 69–88.

influenced' prediction ability. How will Ofqual's statistical standardisation process will deal with such inhomogeneity?

- Dhillon (2005) also suggested that social background might affect teacher-prediction accuracy. There are other pertinent factors such as ethnicity, gender, subject of study and a whole host of others. It is hard to see how a centre-based standardisation approach can satisfactorily deal with all of these fairly. On the other hand, it is also hard to see how a student-centred standardisation could be formulated and executed, certainly in the short time available. So, one has to sympathise deeply with Ofqual.

2. Rankings

- The use of student rankings by teachers within centres is worrying. The main problem I have with rankings is their uncertainty, especially rankings in the middle. For example,

> *"universities, schools and hospitals are regularly ranked in a variety of contexts, the results of which typically generate interest and can often drive policy decisions. In many of these situations, a given ranking can carry a high degree of uncertainty, with this effect particularly pronounced in high-dimensional cases; that is, where there are very many populations or institutions to be ranked."*
> Hall, P. and Miller, H. (2010) Modeling the variability of rankings. *Annals of Statistics*, 38, 2652.

The basic problem is that we usually know which individuals are at the top or bottom of a set of rankings and feel confident about that, but often we have great uncertainty about rankings in the middle:

> *"For example, in the THE-QS university rankings, Harvard University has ranked first for each of the years 2005–2008, while New York University's rankings are 56, 43, 49 and 40 ...we can reinterpret this behaviour as a tendency to obtain correct rankings at extremes, but not otherwise."*
> Hall and Miller (2010), p. 2653.

This phenomenon also applies to teacher-predicted grades as Dhillon (2005), page 79, notes

> *"there is evidence from previous research that the easiest grades to predict are those representing extremes of ability"*

Ofqual has asked exam centres for student rankings, but nothing about their uncertainty.

- Ofqual's Heads of Centre Guidance[3] is clear that for the statistical standardisation *"to be as fair as possible, it is important that the rank order of the students in each subject is as accurate as possible."* (page 8). In light of the evidence above about the uncertainty of rankings in the middle, is it even possible that such rankings can be accurate?

[3] Summer 2020 grades for GCSE, AS and A level, Extended Project Qualification and Advanced Extension Award in maths. Information for Heads of Centre, Heads of Department/subject leads and teachers in the submission of centre assessment grades. Updated 22 May 2020. Ofqual/20/6614/2

Ofqual admit that *"[ranking] will be challenging for some centres and in some subjects and in the current circumstances"* and that for large entry subjects, *"with many teachers, we recognise that this will be challenging"*, both page 8. Since we know the rankings are challenging and highly uncertain, how are Ofqual using this fact in their statistical standardisation?

- Why are middle rankings worrying? The answer depends on what they are being used for and we do not know that yet. One possibility is that Ofqual's standardisation process might use rankings to decide which students are moved up or down a grade. For example, if the statistical standardisation process says that an exam centre has too many students at grade B and that some of those students' grades need to be lowered. Which students will be chosen? Well, possibly the ones that have the smallest rank amongst all students currently holding grade B in that centre. However, the rankings in the middle are very uncertain, so you cannot be sure that you will be moving down the right students. This could be unfair for individuals.

- What else might ranking uncertainty depend on? Possibly, the type of exam centre. For example, Dhillon (2005) provides evidence that FE establishments had a harder job predicting grades and presumably this extends to rankings. Does ranking uncertainty also depend on other factors such as the subject being examined, social background, gender, ethnicity? Ranking uncertainty is possibly greater in some of these cases, which, following on from the previous point, could be even more unfair. More evidence is required.

- Ranking uncertainty will, unfortunately, almost certainly depend on different centres using different methods. The Ofqual Heads of Centre Guidance curiously encourages this *"There are a number of ways in which this could be done"*, page 8 and, sadly, *"it has not been possible to provide national training to school and college staff to standardise these judgements"*, page 9. In the section on *"The importance of objectivity in the Guidance"*, page 9, we see:

  > *"It is important that the centre's grading and ranking judgements are objective; they should only take account of existing records and available evidence of a student's knowledge, skills and abilities in relation to the subject. This evidence should inform teachers' professional judgements about each student's likely performance at the time of the exam. Other factors should not affect this judgement, including characteristics protected under equalities legislation such as a student's sex, race, religion/belief, disability status, gender reassignment or sexual orientation. Similarly, judgements should not be affected by a student's behaviour (both good and poor), character, appearance or social background, or the performance of their siblings."*

  Nobody could disagree with this statement. However, I question its statistical feasibility. Those charged with ranking students will know that there are some groups of students where they cannot discriminate, and/or they are very uncertain about how they compare, and this could be due to a variety of reasons.

- Ofqual touch on these problems with statements such as *"If 2 or more students are almost indistinguishable in terms of their subject performance (and are therefore judged likely to get the same grade) then it may be very difficult to put them into a rank order. However, exam boards will need a single rank order for all students. Tied ranks (that is, giving 2 students position 1) will not be allowed*

and will mean the submission is rejected by the exam board.", page 8 of the Guidance. This seems wrong. Giving two students a tied rank is saying that you cannot discriminate between them. Suppose an exam centre is then forced by the process to give two different arbitrary rankings to students it believes are tied. Then, after the statistical standardisation process suppose one of those students is moved down a grade. This seems unfair, since the ranking was ultimately arbitrary and unnaturally forced by the process.

- The Head of Centre has to include the following in their declaration *"I confirm that these centre assessment grades, and the rank order of students have been checked for accuracy"*, page 17. Given the potential uncertainty in rankings, I do not see how anyone could sign such a statement unless Heads of Centre are internally assessing ranking uncertainty well, in the same way in every Centre. Is this happening? I doubt it.

- In my view, the introduction of rankings fundamentally changes the nature of the UK assessment process. This might be unavoidable due to COVID. Previously, a student would sit an exam, obtain a mark and then this would be assigned a grade by exam boards after establishing grade boundaries. Hence, a student's final grade would depend on what other students achieved, but only loosely through the entire population of students doing that exam. In the proposed post-COVID system, since rankings are collected within an exam centre (and, as far as I know, not being merged with those from other centres) students are being assessed with respect to the performance of their peers in their local exam centre. These are meant to be national exams. We have described a student-focused scenario of what might happen in the annex, point 2.
- So far, I have drawn attention to ranking uncertainty. It might have been preferable for Ofqual to have devised a system where teachers could have expressed uncertainty about their rankings or grades. For rankings, this might have been about reporting groups of tied students. Another possibility might have been not to report a single grade for a student, but a range of grades with associated estimates of the chances of obtaining those grades. See point 3 in the annex for an example of how this might be done. There are other possible methods for predicting grades and expressing uncertainty. The UK has world experts in eliciting uncertainty (not me) that could be asked to suggest practical ideas.

Other Matters and the Future

- *"we have decided not to provide for appeals in respect of the operation or outcome of the statistical standardisation model."*, page 17 of the Guidance. This seems wrong, especially given that nobody knows what the process is until after the grades are released. What if the process is seen to be manifestly unfair or wrong? (Another distasteful aspect of rankings, is that your ranking was changed after appeal, then someone else's ranking changes too, and they could be relegated!)

- It is understandable that it is difficult for Ofqual to share its detailed internal deliberations on the development of its statistical standardisation and wider grade awarding processes. However, it could share them with other parts of Government, for example the Office for National Statistics and/or other parts of the Government Statistical Service for review and comment. These organisations employ leading statisticians and could maintain confidentiality. Has this happened?

- We would advocate that an urgent programme of research, with rigorous high-level statistical involvement and review, be conducted into developing fair systems for producing grades in these

circumstances. Nothing should be ruled out from consideration: this might include use of other data sources, but also admit contributions from online assessment. Such research might be funded by the Government, Ofqual or the Research Councils. We do not know what the future holds, and it would be good to be prepared.

Annex

1. Correlations on grades. Why should caution should be exercised when considering correlations calculated on grades? To use correlation, the grades have to be turned into numbers. Dhillon uses A=5, B=4, C=3, D=2, E=1, U=0, but the choice is arbitrary (and referred to by Dhillon and others, rather worryingly, as a 'plastic interval scale'). Choosing other equally reasonable scales can affect the correlation markedly and, hence, our understanding of how effective teacher-predicted grades are for predicting 'true' grades. In other words, correlations depend on arbitrary grade-to-number mappings, and not necessarily estimates of the intrinsic association between 'true' and teacher-predicted grades.
2. Using Rankings. One can imagine several unattractive possibilities of using rankings. Suppose a bright student had not been working particularly hard (which could be due to social or personal pressures) compared to their peers and their exam centre ranks them lower than other students who have worked consistently hard. Then, suppose that centre over-predicts grades overall and subsequently the statistical standardisation process demands that grades at that centre need to be reduced.
   Pre-COVID, grade judgements would be made with reference to grade boundaries, set nationally. As Ofqual's blog states on grade boundaries "*We, and the exam boards, will have the full national picture*", 3rd February 2017. So, the bright student above might have their grade 'reduced', but only because of the nationally set grade boundaries (of course, there will always be exceptions in unusual cases).

   Post-COVID, grade judgements might be augmented by reference to their local peer group's rankings. So, the bright student's grade will depend on grade boundaries as previously, but also, presumably rankings of their peers (otherwise, why collect rankings?) This seem quite a change.

3. Expressing Uncertainty Example. For example, 'I expect this student to obtain a B grade, with probability 60%, but there is a 30% chance they would get an A, and 10% that they would get a C'. So, the predicted grade would be B, but an exam board could use the extra information if grades needed to be changed. From the teacher's point of view, they have been permitted to express their uncertainty. From the student's point of view there was always the chance that they would obtain a C. This seems better than a teacher predicting a B (only) and then getting a C.

July 2020