ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

12 Errol Street
London EC1Y 8LX
email: education@rss.org.uk
www.rss.org.uk

Mr Ed Humpherson
Director General for Regulation
Office for Statistics Regulation
1 Drummond Gate, London
SW1V 2QQ

14th August 2020

Dear Ed,

We are writing as President and as Vice President, Education and Statistical Literacy of the Royal Statistical Society (RSS) to ask formally that the Office for Statistics Regulation (OSR) conduct a measured review of the statistical models that qualifications regulators across the UK put in place, and the process by which they did so, in order to carry out a statistical adjustment of 2020 predicted exam results arising from the fact that there have been no national examinations due to Covid-19.

The RSS has of course seen your statement of 12 August that OSR would not undertake a review "of the implications of the model for individual results". We want to be clear that the RSS is not seeking that. Given where we are in all the four nations of the United Kingdom, and the tight timetable for appeals and so on, we do not believe that would serve either individuals' interests or the wider public good. The issues of individual grades, appeals and admissions to higher and further education will now be dealt with in other ways.

But the RSS does believe it is essential that a formal review is carried out to address two issues which we have raised since our earliest engagement with Ofqual; we believe our concerns about these have now been vindicated. Before we summarise them, it may be appropriate for us to say why we believe that OSR does indeed have a responsibility to consider this matter.

We believe that a review is essential to address the issue of the extent to which the qualifications regulators did indeed adhere to their obligation to serve the public good. As you have acknowledged, they have a duty to act in accordance with the principles set out in the Code of Practice for Statistics, to achieve 'trustworthiness, quality and value'. We ask in particular whether the models and processes adopted by the qualification regulators did in fact achieve quality and trustworthiness. This is linked to the two issues the RSS has raised since April.

Quality

From the outset, the RSS has expressed sympathy with the challenges of establishing an effective grade-estimation procedure, not least in the time-scale required. We raised various technical questions that we believed any such estimation process would have to consider, given the data available, in order to pass any quality threshold. These can be summarised as:

- The likelihood of systematic upward *bias* (in a statistical sense) in teacher-assessed grades
- The *uncertainty* that was likely to attach to rankings, especially for middle-ranked students
- The *variability* in performance by 'exam centres' (primarily schools and colleges) and whether these were stable enough to bear the statistical weight put upon them.

These are of course complex issues, and we had no privileged access to data when we raised them. We were clear from the outset that we could understand why one concern might be that simply awarding teacher-assessed grades would be biased upwards and might be unfair not only to different year cohorts of students but to those (including higher and education providers and employers) who would use grades in their decision-making. But we noted the importance of considering uncertainty in rankings, especially for those not at the top or bottom of exam centre rankings, and of considering whether variability in results of exam centre distributions of grade results (not just their averages or medians but the features of the distribution) might be a problem, and might also vary between different types of school/college. This might, in turn, suggest a variety of possible uses of individual-level data about prior achievements, not just to assess exam-centre performance, but to inform statistical adjustments in a context of uncertainty.

We should note that we also understood why the qualification bodies could not produce a detailed algorithm before exam centres submitted their data (teacher-assessed grades and rankings). This is so not only for the reasons that you note in your statement of 12 August, but also because there would not have been the empirical data to hand to test and compare various possible models of statistical adjustment.

However, even after our initial examination of the algorithm used by Ofqual (now published), it is not clear to us that these issues have been taken into account, that they could not or should not have been, and that doing so might have resulted in statistically-adjusted grades that gave more weight to individual students' performance, and allowed more clearly for a degree of uncertainty. This might have been at the expense of a somewhat larger uplift in historic trends in exam performance but as we discuss below, this bears directly on the issue of 'trustworthiness' and the relative lack of transparency in the qualifications regulators' approach.

Trustworthiness

One issue underpinning trustworthiness of statistics is their quality and accuracy, which is why we have summarised some of our technical concerns. But another element in trustworthiness is the transparency with which the statistics have been set out and considered, and the extent to which they meet public need. On this ground too, we have concerns about the approach of the qualification regulators, which we have expressed increasingly clearly throughout the months since the decision was taken to cancel exams.

First, your statement of 12th August mentions Ofqual's technical advisory group. The RSS welcomed the formation of such a group (though an announcement that it existed and who its members were was set out only after the initial, and main, Ofqual consultation). We did however have concerns that there were not enough independent external members (who were neither government employees or current or former employees of the qualification regulators). In a letter to Ofqual (and subsequent emails) we suggested that the RSS could nominate two distinguished Fellows who might have relevant statistical expertise. We eventually heard from Ofqual that they could consider these two Fellows, but only with a non-disclosure agreement that gave us real concern. We understood that members of such a group should not give a running commentary in any way, nor divulge any confidential information about exam centres, schools, or the different models being tested – and we wrote back clearly to Ofqual to this effect. But the proposed confidentiality agreement would, on our reading, have precluded these Fellows (who were suggested precisely because of their relevant statistical expertise, and lack of ties to qualification regulators or exam-awarding bodies) from commenting in any way on the final choice of the model for some years after this year's results were released. We set out our concerns about the terms of

the proposed non-disclosure agreement, and restated our willingness to help if a more suitable agreement could be reached. In the end, we did not get an official response to those questions, and our offer to help was not taken up.

We believe this calls into the question one element in the transparency of the process adopted by the qualifications regulators. We would note too that we are not alone in this. It was only after we failed to hear back from Ofqual that we prepared our submission to the House of Commons Education Select Committee. Again, we restated our view that these were complex issues, that difficult judgements would have to be made, that we had offered to help, and that some degree of transparency in the trade-offs and judgements made in the selection of the final model would be essential to public confidence (the 'trustworthiness') of the final statistical model chosen for grade adjustment. In its report, the Select Committee cited our evidence and itself called for greater transparency.

In the end, the only information about the statistical adjustment that was released before Scottish exam results were announced was a general, verbal description of the model Ofqual proposed to use. There were no statistical details, and no clear discussion of the trade-offs or judgements involved. The citation of evidence was, as Guy Nason has shown, thin and, in our view, inadequate. There was no real clarity that the statistical adjustment model being proposed privileged keeping within a percentage point or two of prior national grade distributions, and treating the rankings of individual students as sacrosanct, with no measurement error, and relying on individual students' prior achievements only as part of judging the historical results achieved by particular exam centres, which were also treated as relatively fixed for most types of exam centres, rather than informing individuals' statistically-adjusted grades. The information did not set out the planned approach with what we believe would be the minimum requirements for real 'transparency': the proposed statistical approach and the options considered; the evidence backing that up, including about uncertainty and variability; and a clear justification of the admittedly-difficult choices and trade-offs that would have to be made.

At that stage, we do not believe that any 'gaming' of the system could have occurred. We are aware of the time pressure under which the qualifications regulators were operating. But we believe that they could and should have set out alternative models with a clearer indication of the advantages and disadvantages – and more importantly, the *judgements* that underpinned their choice for wider discussion.

That we were not alone in being concerned about what was meant by the information released by Ofqual at that time is supported by the large volume of queries that the RSS fielded, mainly from specialist educational journalists, about what this meant the model would be, even before the release of the statistically-adjusted exam results from Scotland.

We would stress that none of these observations are a product of hindsight on the part of the RSS – we have been consistent in setting out our statistical concerns and our observations about the need for more transparency.

It may be helpful to end with a statement about why the principle of transparency matters in underpinning the trustworthiness of statistics.

The use of statistics for public good is based only partly on technical statistical issues. Some statistics are technically bad, wrong or worse than others because of the way that data are gathered, or the statistical modelling that takes place. But in many cases, statistics or statistical models are inadequate for the weight being put on them in decision-making, or embed various other judgements that need to be clear. In this case, there are issues about how much 'grade inflation' to allow, and how to be 'fair' to individual students whose rankings may be uncertain or who are in exam centres whose performance may be less fixed over time than the modelling seems to rely upon. So while we continue to have concerns about various technical decisions

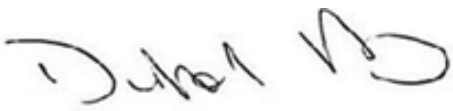
made by the qualification regulators, we also believe that having an more open discussion about this *well before individual results* were announced would have resulted in more trust in, and more trustworthy, statistical choices, in part because there would have been greater understanding of the underlying principles being applied and more detailed justifications of them. That is particularly important given that judgements about what is 'fair' have featured so widely in Ofqual's statements and in other commentary.

'Fairness' is not of course a statistical concept. Different and reasonable people will have different judgements about what is 'fair', both in general and about this particular issue. But real transparency would have enabled a deeper, earlier and better public discussion not only about the technical issues we have raised, but would have allowed that to be divorced from the 'strong interest' you mention in your statement of August 12. This is not because those interests are wrong, or unimportant. But a statistical procedure should be capable of being judged as 'fair' or 'reasonable' in advance of its being used or knowing which individuals may be affected. This is one reason the RSS puts so much weight on transparency in its Data Manifesto. We do not believe that the development of the statistical adjustment methodology has been transparent enough to meet our concerns about statistical quality or the need for greater involvement of knowledgeable external experts. We are sure that it has not been sufficiently transparent to meet the aim of being trustworthy in the broader sense.

These issues would be worthy of consideration even if we could be sure that the cancellation of exams were a one-off occurrence. But none of us can be certain that the UK will not face similar issues in future. We can, however, be sure that the broader question of transparency in the use of algorithms by public bodies, and its importance to the quality and trustworthiness of statistics, will recur in other areas. An OSR review would seem essential both to address the questions that have arisen this year and to set a benchmark to ensure they do not happen in the future – in this domain or in others.

We look forward to your reply.

Yours sincerely,



Professor Deborah Ashby OBE FMedSci

President of the Royal Statistical Society

Sharon Witherspoon MBE FAcSS

Vice-President of the Royal Statistical Society, Education and Statistical Literacy