A cutting re-mark

In mid-August, an algorithm spat out "synthetic" exam grades so absurd and distressing to students that they had to be ditched just four days later. But what exactly went wrong? **Dr Tim Paulden** explains

Perhaps the most curious aspect of this summer's exam grading debacle in England (see News, page 2) was the *Black Mirror*-esque vibe that seemed to permeate the national discourse around the offending computer code. Prime Minister Boris Johnson dubbed it the "mutant algorithm" – conjuring up images of a twisted silicon supervillain handing down life-changing judgements. The reality, though, was rather more pedestrian: the algorithm developed by England's exam regulator, Ofqual, contained no elements of machine learning, or indeed any artificial intelligence – it was simply a sequence of hand-coded, procedural steps designed by an indisputably human team.

When Ofqual's synthetic A-level grades were released on 13 August, there was immediate uproar. Almost 40% of the predicted grades submitted by schools had been revised downwards (bit.ly/2RnmhZA). Within hours, school and college heads came forward to protest that the synthetic grades were excessively volatile and, in many cases, nonsensical (bit.ly/3bUwmGW). Four days later, under increasing public pressure, Ofqual rescinded the synthetic grades and reverted to teacher-assessed grades – despite these being undeniably generous on an aggregate level (bit.ly/2RqfNcE).

What went wrong?

Algorithm is gonna get you

Ofqual's hefty 319-page report (bit.ly/32qiW2u) and implementation document (bit.ly/2RlHWBw) explain how the algorithm was meant to work:

(1) A synthetic 2020 grade distribution would first be generated, for each subject in each school, by taking the school's historical A-level grade profile for that subject, and adjusting it up or down according to whether the school's 2020 cohort had stronger or weaker prior grades at GCSE level than previous cohorts. (GCSEs are the qualification taken at age 16, prior to A levels.)

(2) Students would then be allocated grades from this synthetic distribution, working down the teacher's ranking list, and an implied "mark" imputed for each student.

(3) Finally, grade "breakpoints" would be moderated on a national level – potentially nudging some students between grades – to achieve the desired national grade distribution. The Department for Education had made clear it did not want to see grade inflation in England, meaning that the overall grade distribution would need to be broadly in line with previous years.

On the surface, the above steps seem fairly intuitive. However, when reading through Ofqual's report, several questionable aspects of the methodology quickly become apparent. The six central points of controversy may be summarised as follows:

Class-size disparity

As has been widely reported (bit.ly/3bX7yyj and bit.ly/32tXfOW), Ofqual's algorithm accepted the teacher-assessed grades – referred to by Ofqual as "centre-assessed grades" (CAGs) – as being correct for very small classes of up to 5 (judged, to be precise, using the harmonic mean of the current class size and historical class size; bit.ly/3htnHg2), and granted the CAGs some weighting for small class sizes of up to 15 (similarly judged). For larger classes, however, the CAGs were assigned zero weight, with only the teachers' rankings of students being used. This disparity is highly controversial, since CAGs tend to be overly generous, and smaller class sizes are more common in independent schools. It appears that, to avoid grade inflation overall, schools with medium or large classes "took the hit" to counterbalance the generosity towards smaller classes.

Misevaluation of outlier students

For most students (except those in small classes), Ofqual's synthetic grade calculation did not directly use either the CAG value or a student's individual GCSE performance – only their position in the teachers' rankings. This means that an exceptional student at a school without a history of high A-level grades might be denied the A* confidently predicted by their teachers, despite achieving top grades at GCSE. Disregarding the CAGs potentially opens up Ofqual to the criticism that their algorithm attempted to *predict* students' grades – a monstrously difficult task – rather than genuinely *moderating* teachers' judgements.

No performance trajectory

A school's historic A-level performance in a subject was calculated using a three-year average, without accounting for any upward or downward trajectory in performance. This is a particularly vital consideration for new and improving schools, and without it, such schools find themselves doubly punished by Ofqual's algorithm: not only is the performance trend neglected, but the three-year average (and, thus, the 2020 prediction) for an improving school will sit *below* the school's 2019 performance, and vice versa for a school trending downwards. The method therefore appears to "level down" by systematically pulling the predictions towards the centre.

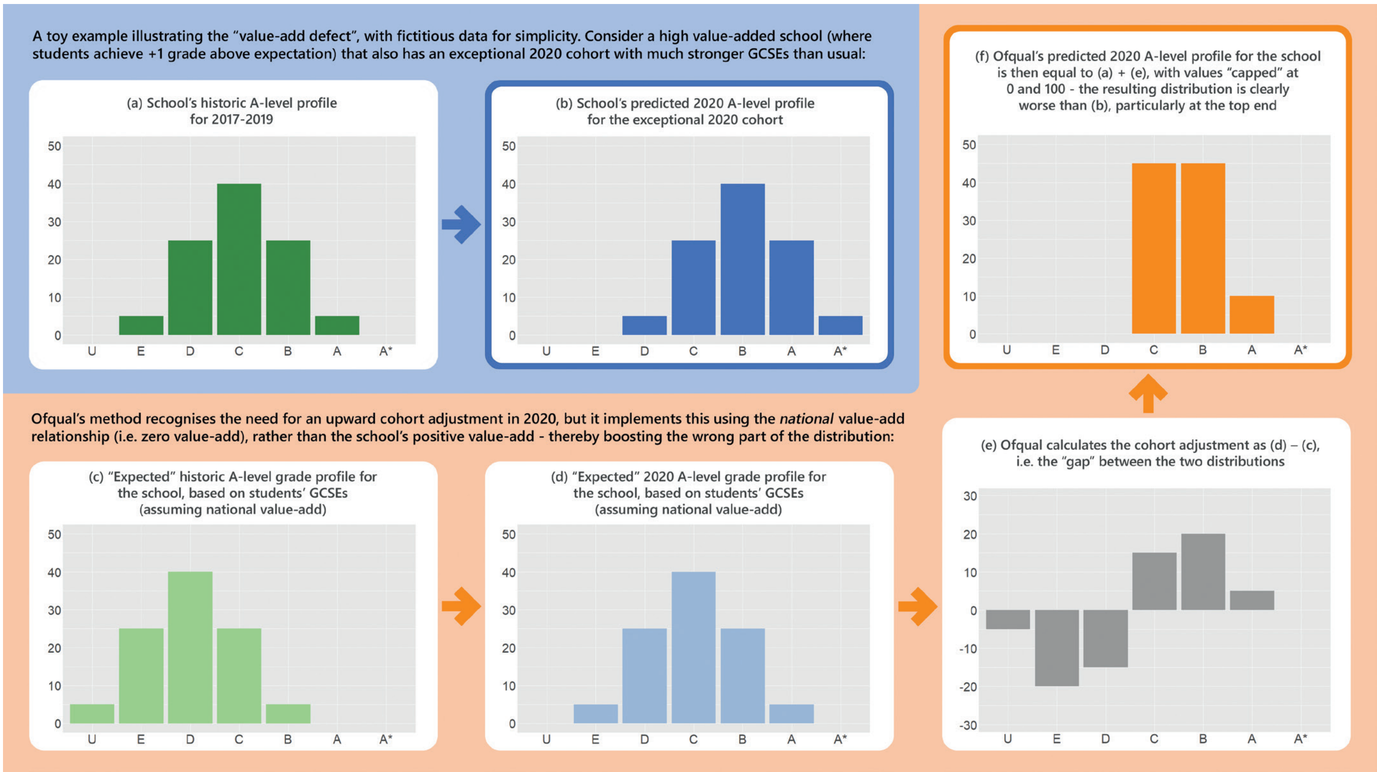
The "value-add defect"

In education, the concept of "value-add" quantifies how much a school improves students' grades between GCSE and A level, compared to the national average. Ofqual's method of adjusting a school's synthetic grade distribution to account for the strength or weakness of the 2020 cohort (as described in (1) above) suffered from a pernicious problem we might call the "value-add defect" – namely, the cohort adjustment was evaluated using the *national* value-add relationship (i.e. zero value-add) rather than the school's own value-add.

The example in the diagram illustrates the dramatic effect this defect can have on a hypothetical school with a high value-add of +1 (that is, students have historically achieved A-level grades that are – on average – one grade above those predicted by their GCSEs), and a 2020 cohort that is stronger than usual. It turns out that the synthetic grade distribution generated by Ofqual's algorithm (plot (f) in the diagram) has an implied value-add of just +0.65 – well below the school's historical value-add of +1. In other words, the algorithm is once again "levelling down" towards the average. There is considerable

© 2020 The Royal Statistical Society

Dr Tim Paulden is the innovation and development manager at ATASS Sports, and a long-standing governor at Exeter Mathematics School.



A toy example illustrating the “value-add defect”, with fictitious data for simplicity. Consider a high value-added school (where students achieve +1 grade above expectation) that also has an exceptional 2020 cohort with much stronger GCSEs than usual:

(a) School’s historic A-level profile for 2017-2019

(b) School’s predicted 2020 A-level profile for the exceptional 2020 cohort

(f) Ofqual’s predicted 2020 A-level profile for the school is then equal to (a) + (e), with values “capped” at 0 and 100 - the resulting distribution is clearly worse than (b), particularly at the top end

Ofqual’s method recognises the need for an upward cohort adjustment in 2020, but it implements this using the *national* value-add relationship (i.e. zero value-add), rather than the school’s positive value-add - thereby boosting the wrong part of the distribution:

(c) “Expected” historic A-level grade profile for the school, based on students’ GCSEs (assuming national value-add)

(d) “Expected” 2020 A-level grade profile for the school, based on students’ GCSEs (assuming national value-add)

(e) Ofqual calculates the cohort adjustment as (d) – (c), i.e. the “gap” between the two distributions

Illustrating the “value-add defect”: In the hypothetical school shown in the diagram above, students have historically achieved a C-grade average at A level (a), but were expected to achieve a D-grade average at A level based on their GCSEs, assuming zero value-add (c). The school’s 2020 cohort is much stronger than in previous years: their expected A-level results, based on GCSEs, are a C-grade average, assuming zero value-add (d). When the school’s +1 grade value-add is factored in, the school can reasonably predict that the 2020 cohort will achieve a B-grade average (b). Ofqual’s approach to the cohort adjustment is quite different, however: it calculates the bar-by-bar differences between the 2020 cohort’s grade expectations (d) and the historic grade expectations of the school’s previous cohorts (c), both of which are based only on GCSE results and assume zero value-add. The difference between these two distributions (e) is then added to the school’s historical A-level profile (a) to create the school’s synthetic grade distribution for 2020 (f). It is immediately evident that the distribution has been “boosted” in the wrong place, and is far too mean at the top end compared to (b). (See bit.ly/3ioNlUl for an interactive visualisation by Tom Haines.)

evidence that high value-add schools were systematically short-changed by Ofqual’s algorithm and that this “value-add defect” was an important contributor. The unintuitive, non-standard manner in which the grade distributions are manipulated, chopped up, and capped also raises concerns, as noted by Tom Haines at the University of Bath (bit.ly/3hpf1ad).

The U-grade glitch

When allocating grades based on a school’s synthetic grade distribution and its teacher ranking of students (as described in (2) above), it appears that the algorithm’s rule-set contained a subtle flaw whereby the bottom student would often be assigned a U grade, even when the probability associated with a U grade was tiny – representing less than a single student (bit.ly/2RgtW8K). It seems this problem can arise even for schools with no history of low grades, due to the cohort adjustment discussed earlier – if a school has a relatively weak 2020 cohort, a small U-grade “slice” may propagate into the synthetic grade distribution.

Faulty model testing

Finally, as highlighted recently by Haines (bbc.in/2Rrc12O), Ofqual’s procedures for “back-testing” the model on historical data were faulty because the presumed “teacher rankings” for previous years were based on actual final grades. Model performance metrics from this testing phase were therefore over-optimistic, as the algorithm was essentially “peeking” at the true rank order in which students had finished, and using that same ranking to assign synthetic grades – significantly reducing the potential for error.

Regulators, mount up

Though it is not possible to quantify the relative impact of each factor without access to Ofqual’s code or data, the directionality of each one – and the schools or students most affected – should be evident.

At the time of writing, the Office for Statistics Regulation (OSR) is reviewing Ofqual’s algorithm, and other inquiries are expected. A group of governors from Exeter Mathematics School (myself included) will be writing to the OSR to share the concerns enumerated above, and to propose that the review includes concrete recommendations – particularly around model testing and validation – to avoid similar issues arising in future.

It is imperative that the public can trust the algorithms and models that impact our lives – and few applications carry greater consequence than the algorithmic setting of exam grades. ■