ARTICLE

Exam algorithms: some lessons

Algorithms are powerful tools for analysing large amounts of data and producing predictions. The uproar following the use of an algorithm to assess UK exam results has damaged perceptions over their reliability. How might public confidence in them be restored?

Context

One of the most significant policy decisions made by the UK Government in response to the COVID-19 pandemic has been to close all schools and cancel all GCSE and A level examinations in summer 2020. This policy decision reflects a trade-off between the public health benefits of potentially slowing the spread of the virus, and the costs to the economy and to society, including disruption to the education system. To manage this disruption, the UK Government also directed the UK's Office of Qualifications and Examinations Regulation ("Ofqual") to produce alternative arrangements for grading students. It specified that "*students should be issued with calculated results based on their exam centres' judgements of their ability in the relevant subjects, supplemented by a range of other evidence*", and that "*Ofqual should ensure, as far as is possible, that qualification standards are maintained and the distribution of grades follows a similar profile to that in previous years*."[1] On 15 April 2020, Ofqual launched a public consultation around how to produce such calculated results.[2] On 13 August 2020, in parallel with the A level results, Ofqual published its final methodology in a report of over 300 pages, containing detailed, technical research and analysis.[3]

Ofqual's algorithm

In simple terms, Ofqual's methodology employs a set of rules (i.e. an 'algorithm') to predict the grades that each student would have obtained had they sat their exams as per normal, using information about:[4]

- the *school's* historical exam performance;
- the current students' past exam performance (for example at GCSE level);
- the teachers' ranking of students for each subject; and
- the teachers' grade predictions (referred to as centre assessment grades, or "*CAG*").

The A level results and methodology were met with widespread public outcry about the outcomes, which have been perceived as unfair, discriminatory, and arbitrary. Despite Ofqual's own analysis concluding that there is "*no evidence that this year's process of awarding grades has introduced bias*", and "*changes in outcomes for students with different protected characteristics and from different socio-economic backgrounds are similar to those seen between 2018 and 2019*",[5] there have been:

[1] Directive to Ofqual, Rt Hon Gavin Williamson, 31 March 2020.

[2] Exceptional arrangements for exam grading and assessment in 2020, Ofqual, 16 June 2020.

[3] Awarding GCSE, AS, A level, advances extension awards and extended project qualifications in summer 2020: interim report, Ofqual, 13 August 2020.

[4] Awarding GCSE, AS, A level, advances extension awards and extended project qualifications in summer 2020: interim report, Ofqual, 13 August 2020, section 8.

[5] Awarding GCSE, AS, A level, advances extension awards and extended project qualifications in summer 2020: interim report, Ofqual, 13 August 2020, section 10 on Student-level equalities analysis.

- **numerous anecdotal reports of high-achieving students being awarded low grades** – significantly below their own expectations, their teachers' predictions, and the requirements of their offers to study their preferred courses at prestigious universities;[6]
- **observations that students in different socio-economic groups have experienced different outcomes** – for example, independent schools (i.e. those which charge fees) experienced a 4.7% point increase in the proportion of students being awarded A or A* grades, which is more than twice the figure for secondary comprehensive schools (2%), and more than 15 times the figure for sixth form colleges (0.3%);[7]
- **criticisms about particular aspects of the design of the algorithm** – for example, Ofqual's algorithm explicitly awards teachers' predicted grades (which it acknowledges tend to be overoptimistic) to students in the smallest classes. Because schools in wealthier areas tend to have smaller class sizes, this design feature of the algorithm may have the unintended effect of favouring students from wealthier backgrounds, and penalising those from poorer backgrounds (for whom the teachers' predictions either receive less weight, or are ignored altogether); and
- other comments by statisticians and data experts around various technical issues with the design, implementation and testing of the algorithm, and the threat of legal action against Ofqual, in line with the pre-action protocol for a Judicial Review.[8]

Within days, Ofqual decided to abandon the algorithm entirely, resorted to awarding students their teacher predicted grades, and issued an apology.[9]

Lessons for the proper design and use of algorithms

This controversy has highlighted a number of important questions around the proper use of algorithms, and the assessment of any bias in the outcomes produced by those algorithms. Although Ofqual's algorithm has now been withdrawn, these questions remain pertinent to the use of algorithms more generally, and will be of increasing importance in the future, as the analysis of ever larger and richer datasets continues to inform the design and evaluation of policy decisions and business strategy. There is little question that the reputation of algorithms in the public mind has taken a substantial hit: this matters because algorithms are an increasing fact of life and directly impact all of us in many different ways, from the presentation of media choices to consumers, to the provision of credit, the pricing of goods and services, the diagnosis of disease and so on.

Here are some of the lessons that are already emerging:

Lesson 1: an algorithm should only be described as biased if it systematically and unfairly discriminates against certain individuals or groups of individuals in favour of others[10]

This is important, because an accusation of bias has serious (and potentially legal) implications. From a statistician's perspective, for an algorithm to be biased, it must *systemically* and *unfairly* treat one group of individuals differently to others (i.e. discriminate in the technical sense of the word, instead of the legal sense). For example, if Ofqual's algorithm excessively moderated all A level grades down to temper grade inflation, this systematic adjustment would not be *differentially* unfair to any particular set of students, so it would not be biased against any particular group. Similarly, if Ofqual's algorithm resulted in some high-achieving students being awarded lower grades than they would otherwise have achieved, this would of course be an *unfair* outcome for those particular students – but unless these errors were systematically affecting particular groups of students, the algorithm would not be biased. It is inevitable that some unfair (even if unbiased) outcomes will occur, and this is why a good appeals process is required. This applies not only to exam results, but to algorithms in general.

> *"From a statistician's perspective, for an algorithm to be biased, it must systemically and unfairly treat one group of individuals differently to others."*

[6] A-levels and GCSEs: Student tells minister 'you've ruined my life', BBC, 15 August 2020.

[7] Awarding GCSE, AS, A level, advances extension awards and extended project qualifications in summer 2020: interim report, Ofqual, 13 August 2020, Table 9.10 Outcomes by centre type at grade A and above (2018 – 2020) (percentage).

[8] Law firm threatening legal action over A-level grades, Leigh Day, 16 August 2020.

[9] Statement from Roger Taylor, Chair, Ofqual, 17 August 2020.

[10] Friedman, B. & Nissenbaum, H. (July 1996), Bias in Computer Systems, ACM Transactions on Information Systems, Volume 14, Number 3, pages 330-347.

Lesson 2: even with the best of intentions and good execution, algorithms can (and do) make errors

In general, the accuracy of a predictive algorithm can be assessed by comparing its prediction against the actual outcomes. In some cases, this is easy to do: algorithms used to predict share prices movements can be assessed by comparison to actual share price movements. In other cases, it may be impossible: for example, students will never actually be able to sit their summer 2020 exams in the absence of the COVID-19 disruption, so the accuracy of Ofqual's A level algorithm at predicting such grades will never be known. In such circumstances, the next best alternative is to test the algorithm's accuracy using historical data. Ofqual used its algorithm to predict 2019 results for some subjects using data prior to 2019. It compared those predictions to the actual 2019 results, and found that approximately 50% to 75% of grades were correctly predicted within one grade, depending on the subject.[11] This implies that 25% to 50% of grades were incorrectly predicted, with the error being more than one grade (for example, predicting a C, when the result should have been an A).

Algorithms that produce more accurate results are obviously preferable, but accuracy is not the only consideration: the nature and distribution of the errors is also very important. Ideally, any errors should be small in size and randomly distributed. If there are large errors, or if errors are concentrated among certain groups of individuals, then there may be cause for concern. In the case of Ofqual's A level algorithm, there have been many anecdotal reports of students being awarded grades significantly lower than teachers' predictions (although these are not necessarily errors – i.e. some students might be expected to fail to achieve their teachers' predictions, just as some are likely to over-achieve), and evidence that grade inflation is lower for certain groups of schools (although again, it is not obvious that this result is necessarily an error – for example Ofqual identifies evidence in the literature that teachers tend to over-predict students' actual grades, and that there tends to be more over-prediction for more disadvantaged students).

> *"Algorithms that produce more accurate results are obviously preferable, but accuracy is not the only consideration: the nature and distribution of the errors is also very important."*

Lesson 3: specialist statistical analysis is required to design and assess algorithms

Parts of Ofqual's algorithm are relatively transparent, in that there are explicit rules that determine grades: for example, if the class size is 5 or below, Ofqual assigns the teachers' predicted grade. However, algorithms are not always as such: those built using machine learning methods can be extremely complex, and may be practically impossible to assess on a rule-by rule basis. Instead, formal statistical analysis can be used to assess whether the outcomes of the algorithm (for example, a student being assigned a grade that is below her teachers' prediction) are systematically related to particular characteristics (for example, whether the student belongs to a particular socio-economic group). Ofqual performed some such analysis in its interim report, and concluded that its algorithm introduced no bias – although its analysis has not yet been fully subject to independent scrutiny. On 18 August 2020, the Office for Statistical Regulation announced that it will conduct a review of Ofqual's approach.

Lesson 4: organisations that use algorithms in decision making should expect the results to be scrutinised and prepare accordingly

Although algorithms are powerful tools that can produce eerily accurate predictions (think of the product suggestions that you see on your Amazon account home page, or the predictive keyboard on your smartphone), when they are used in public and high-stakes decision making (such as to determine examination grades for hundreds of thousands of students), they will inevitably be subject to intense scrutiny. Organisations should anticipate this review. They should:

- **prepare clear and transparent communications**, to explain the objectives of the algorithm, how the algorithm works in simple terms, how accurate its predictions are, and what checks have been made to ensure that its outcomes are not biased;

[11] Awarding GCSE, AS, A level, advances extension awards and extended project qualifications in summer 2020: interim report, Figure 7.3 Overall predictive accuracy for the different models for A level biology, French, drama and religious studies, DCP approaches 1, 2, and 3.

- **implement a fair and proportionate appeals process**, in which a human agent examines allegedly erroneous and unfair outcomes, and has the power to overturn those outcomes if appropriate; and
- **retain independent experts** to inspect and test their algorithms, using specialist statistical analysis.

This process will produce outcomes that are more likely to be free of unintended consequences – and that is a good thing, not only for the organisation using the algorithm, but for public confidence in the use of algorithms in general.

MELORIA MESCHI, PH.D
Senior Managing Director
Economic and Financial Consulting
+44 20 3727 1362
meloria.meschi@fticonsulting.com

DAVID EASTWOOD
Senior Managing Director
Economic and Financial Consulting
+44 203 727 1292
david.eastwood@fticonsulting.com

RAVI KANABAR
Senior Director
Economic and Financial Consulting
+44 20 3727 1280
ravi.kanabar@fticonsulting.com

The views expressed in this article are those of the author(s) and not necessarily the views of FTI Consulting, its management, its subsidiaries, its affiliates, or its other professionals. FTI Consulting is an independent global business advisory firm dedicated to helping organisations manage change, mitigate risk and resolve disputes: financial, legal, operational, political & regulatory, reputational and transactional. FTI Consulting professionals, located in all major business centres throughout the world, work closely with clients to anticipate, illuminate and overcome complex business challenges and opportunities. For more information, visit www.fticonsulting.com and connect with us on Twitter (@FTIConsulting), Facebook and LinkedIn. **www.fticonsulting.com**.