Is the algorithm working for us?

Algorithms, qualifications and fairness
Reflections on the qualifications and the pandemic from the former Chair of Ofqual, Roger Taylor
June 2021

centre for **progressive policy**

Introduction

In 2020, Covid closed 90% of the world's schools. In the long list of harms caused by the pandemic, the disruption to the education of a generation will be felt for a long time. One acute aspect of this was the difficulty countries faced in administering examinations and deciding which children were eligible for higher education and other opportunities.

Some countries moved exams online, some delayed them, others held them in socially distanced environments. Governments in the UK and the Republic of Ireland took an unusual – in some eyes the most extreme option[1] – of using other information sources, teacher-assessed grades and statistical forecasts, to predict what young people would have achieved had they taken an exam. The plan did not work.

There have been a number of reviews into what went wrong and there will, no doubt, be more. This paper sets out some personal reflections on the causes of the problems in 2020 together with some tentative views about how we rebuild after the pandemic.

Roger Taylor

Key points

1

The mistakes were made by humans, not machines. The exam grades debacle of 2020 has been blamed on a malfunctioning algorithm.[2] But by blaming the algorithm, we risk missing the most important lessons on mistakes that were made. The problem was not the algorithm, it was what we were trying to do with it: it was human decision making that failed.

2

Algorithms are helping people make fairer decisions on the basis of qualifications. The debacle has given algorithms an undeserved reputation as a mechanism of injustice, but when used well – to interpret exam results, not replace them – they are proving a powerful mechanism for fairer decision making. Employers and university admissions officers are using data-driven systems to understand qualifications in context. In Australia an algorithm was used, with public support, to adjust exam results for lost learning. Algorithmic decision making is a powerful tool to support the interpretation of qualifications and to increase social mobility.

3

Qualifications need to adapt to digital recruitment methods to ensure fair recruitment. There is evidence that general qualifications are becoming less relevant to selection for employment, a trend that reflects the rising power of digital technology but one that risks undermining the value of qualifications and undermining fair selection. As we plan for rebuilding after the pandemic, there is an opportunity to consider how qualifications can adapt to make sure they retain their power in a data-driven age as a way for young people to progress from education into adult life.

4

Data and information should work for everyone, not just government. A theme running through these reflections is the mindset in government and public administration which too often regards official data and information systems as there to help it achieve its objectives. We might make better decisions if instead we viewed them more as public utilities that exist to enable citizens and civil society to achieve their ends.

1 https://blogs.worldbank.org/education/examinations-and-high-stakes-decision-making-era-covid-19

2 For an explanation of how the algorithm worked see: https://youtu.be/EX5STb0qbGI

The 2020 exam debacle: how did it happen?

The exam debacle of 2020 is remarkable for two particular features: the broad consensus in advance that it was the right thing to do; and then, in the event, the overwhelming rejection by the people affected. The consensus crossed party lines: Labour in Wales, SNP in Scotland, Conservatives in England and the Northern Irish administration all supported the approach. Teachers' leaders, universities, schools and colleges also supported the approach. Even students, in advance of the results, could understand why it seemed the sensible thing to do. When a misjudgment happens on this scale it warrants reflection. How could quite so many people be so wide of the mark?

There is little agreement about what exactly went wrong. The fiercest critics look at the execution: the algorithm was not accurate enough or biased, communication needed to be better, wider consultation would have improved both. There are undoubtedly lessons to learn here, but they cannot adequately account for the failure. Variations on the approach were tried by four separate administrations. All failed. Ireland reversed course at the last minute but still found itself in court. A related scheme operated by the International Baccalaureate Organisation was pulled.

If we look at failings in the execution, and in every case there are things that could have been done better, we miss the more important lessons. Nor should we put the blame on the unprecedented circumstances: a once in a century pandemic, limited time, wholesale disruptions. These all contributed to the debacle. But this still leaves open the question as to why so many people, operating in different contexts, made such consistent misjudgements. There was something more fundamentally amiss.

I have worked with data and technology in the public sector for over 20 years, and have consistently observed a mindset that sees data and statistics as tools for government and public authorities to fix their problems, rather than as tools for citizens to address theirs

Errors of judgement are partly a consequence of the pressures of the moment and partly a consequence of the attitudes and behaviours we bring to that moment – the preconceptions that incline us to do one thing rather than another. When there is broad consensus and that consensus proves wrong, we need to look at the common assumptions that took us in the wrong direction.

The Office for Statistical Regulation has suggested that part of the problem was over-optimism about what algorithms can achieve and its review advised all involved to recognise that algorithms may not be the right answer to a problem. This is part of the answer, but it is also true that much was known about the limitations of the approach in advance and it was still regarded as the best thing to do.

To properly understand what went wrong we need instead to look at the way that the problem was framed and understand how ways of thinking about data and public administration led so many people to chose a course of action so unacceptable to the population. I have worked with data and technology in the public sector for over 20 years, and have consistently observed a mindset that sees data and statistics as tools for government and public authorities to fix their problems, rather than as tools for citizens to address theirs. This is not an issue of party politics or individual attitudes, nor is it malicious or corrupt; it is driven only by good intentions to solve problems that affect us all. But it is prevalent and it is corrosive to public trust and effective use of data.

"How could they be so naive..."

I was chair of Ofqual throughout 2020 and left at the end of the year, shortly before the decision was taken to cancel exams this year. I was, at the same time, chair of the newly created Centre for Data Ethics and Innovation, a body set up to advise the government on the ethical use of data.

Throughout 2020 I was working at Ofqual on how to deliver the algorithmically moderated grades and defending the approach publicly. At the same time I was working with the CDEI on our report on algorithmic bias, which pointed out that people are instinctively distrustful of algorithms, much more so than of human systems. The report set out why it is impossible for algorithmic systems to simultaneously meet all the definitions of fairness people care about.[3]

One conclusion you might draw from these findings is that use of algorithmic systems is unlikely to command public support unless the benefits of using it self-evidently outweigh the alternative options.

The tension between these two halves of my life led to some awkward moments. I recall a Zoom meeting with academics from around the world discussing how governments might improve the governance of algorithms. I made a point about public trust, which was endorsed by another participant who shook his head and asked, "How could those people in the UK have been so naive as to imagine that they could use a predictive algorithm to award places at elite universities?"

I don't think he was aware of my involvement. But his question is the right one. How could we have been so naive as to think you could hand out highly prized and contested places on the basis of an estimate of what might have happened if exams had taken place? How could so many people have gone along with a decision regarded by others as hopelessly naive?

The problem with bias and accuracy

Some of the most common accounts of what happened point us in the wrong direction and risk the same mistakes occurring in the future. Bias is the most frequently cited problem. The Ada Lovelace Institute, in a blog at the time, analysed the problem as follows:

> "This model [algorithmic moderation] prioritises avoiding grade inflation, getting the 'right' school-level results and maintaining the distribution shape over the fairness and accuracy of individual results."

They gave the example of bright kids in poorly performing schools being unfairly marked down.[4] The blog goes on to say that you can't really blame Ofqual because these were the instructions of the Secretary of State.

It is wrong to suggest that Ofqual could be excused if it had used a biased algorithm on the grounds it was told to maintain grade standards. Ofqual was very aware that if the algorithm was biased, it would have to scrap the approach the authorities in all four administrations had looked carefully to check this was not the case. The report by the Office of Statistical Regulation recognises this. It says:

> "All the regulators carried out a variety of equality impact analyses... based on the premise that attainment gaps should not widen, and their analyses showed that gaps did not in fact widen."

This is the definition of fairness that Ofqual had agreed in consultation and it was met.

People rightly observed that aspects of the algorithm operated in favour of, or against particular groups. For example, the inability to apply moderation to small classes disproportionately favoured private schools. However, this is very different from saying the algorithm overall favoured private schools. After moderation, the proportion of higher grades going to private schools was lower than before the moderation process.[5]

The same is true for high-performing children in poor-performing schools. There were certainly instances where bright kids in schools with few or none like them in the past, got unfairly marked down and would have had to appeal their grades.[6] But overall the algorithm did not disadvantage this group. The proportion of A* and A grades going to students in more deprived areas was lower with teacher assessed grades than with algorithmically moderated grades.[7]

"Despite this analytical assurance", the Office for Statistical Regulation goes on to say, "there was a perception when results were released that students in lower socio-economic groups were disadvantaged by the way grades were awarded. In our view, this perception was a key cause of the public dissatisfaction." This is true. So why were people not reassured by the answers provided? Or as my colleague put it, why were we all so naive as to think they might be?

No-one thought algorithmically moderated grades would be uncontroversial. Everyone knew it was fraught with risk. There was widespread unease about the chances of the plan working. No-one thought that moderated grades would have the legitimacy of exams.

Accuracy was always the issue. By accuracy, I am not referring here to the 'obviously wrong' results. The candidates who saw large 'inexplicable' changes between their teacher assessed grades and their moderated grades. The route by which these results ended up in the awarded grades in England is one of the more tangled elements of the whole saga. It attracted headlines and undermined trust. It was entirely predictable that this would happen and it is understandable that people struggle to understand how it was allowed to happen. It is an issue that warrants close inquiry (see Appendix 1).

But we will learn the wrong lessons if we think fixing these 'obviously wrong' results would have made things work. This problem was primarily a feature of the English approach. It was handled differently in other countries, but the protests were equally forceful everywhere. This problem affected relatively few people (0.2% of results in England). However, the sense of grievance was felt by far larger numbers of students. To understand this much wider sense of injustice we need to look how the majority of students were affected by algorithmic moderation – those who saw one or more of their results reduced by one grade.

The problem of accuracy in this much larger number of results was known from the outset.[8] Ofqual raised the problem publicly in its consultation documents in the spring and at its summer symposium in June. It explained why lowering grades through moderation would leave many candidates with lower grades than they would have got in an exam, while others would get higher grades. Unfortunately, there was no way of knowing who they were and so there was nothing that could be done about it.[9] It is true that no algorithm can fix this problem. However, it could have been addressed by adopting a different policy.

A place at university is one of the most valuable and life-changing things that society can offer. Children hoping to go to university in 2020 had been working for years towards that goal. They had been told the process by which they would be given the chance to prove their worth. It is a huge ask of somebody to accept that their chances of going to university and of an opportunity that could transform their lives had been taken away on the basis of an uncertain prediction on what might have happened if exams had not been cancelled.

How could we have been so naive as to think that people would accept this? How could we have formed such a strong consensus that it was the right thing to do? Early on there was hopeful talk about the pandemic creating a 'wartime' environment in which people would understand that their lives would be disrupted, that it was nobody's fault, it was simply the inevitable consequences of the pandemic. Maybe there was some truth in that in March, but it was certainly not true by August and queasiness about the whole plan was growing. Despite this, a consensus formed that there was no alternative.

3 The statistical properties that make this true of an algorithmic system are, of course, equally true of human bureaucratic systems. However human organisations usually comprise (or are viewed as) multiple actors making slightly different calculations so attempts to attribute the distribution of outcomes to a single mechanism are tenuous.

4 Using a different definition of fairness, this analysis makes sense which is perhaps what was intended (see below).

5 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/945861/Summer_2020_results_analysis_-_GCSE_AS_and_A_level_171220.pdf

6 For an explanation of why these results were left to be appealed rather than corrected in advance see appendix.

7 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/938869/6713_Student-level_equalities_analyses_for_GCSE_and_A_level.pdf

8 Throughout this discussion, I am treating exam results as the 'correct' results and variation from this as inaccuracy, since the aim of the algorithm was to predict what the outcome of exams would have been. I am putting to one side the broader context in which both exam grades and calculated grades are imperfect estimates of a 'true grade' – a context which raises the issue of legitimacy vs accuracy. I am avoiding the word 'valid' as it might imply a relationship to 'true grades'. Analysis published by Ofqual noted the possibility that the calculated grades in some subjects might be as close or closer to 'true grades' than examined grades would have been. This does not, as 2020 demonstrated, affect their legitimacy. Legitimacy is not well defined in this context. But a legitimate process might be considered as one where the causes of error and variation are understood, can be explained and/or are accepted for whatever complex social reasons. This paper assumes that it can be rational to prefer a legitimate but relatively inaccurate system over a more accurate but illegitimate alternative. So for the purposes of this discussion, I am treating the non-existent 2020 exam grades as the 'correct' and 'legitimate grades'. Inaccuracy refers to degree to which calculated grades failed to accurately mimic exams. There are some further comments on legitimacy at the end of this paper.

9 In its consultation document Ofqual highlighted the fact that both teacher estimates and rank orders would have a significant degree of inaccuracy. The rank orders had an particularly significant impact on calculated grades. The consultation response document that many people would feel they had got the wrong grade and would have to rely on the autumn exams to correct this. It added that there was no way to distinguish between students who grades were wrongly moderated down by the algorithm and those whose grades were correctly moderated down. In the Summer Symposium Ofqual said that many of the 'optimistic' grades that would be reduced by moderation would, in fact, have been correct. But there was no way of knowing which so it was fairer to apply moderation across the board. Lifting moderation increased the chance that someone got a higher grade than you because their teacher was more generous in their grading and, in that sense, is less fair. But it greatly reduced the chance that anyone was awarded grades lower than they would have got in an exam. In that sense, it was fairer. See: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879627/Exceptional_arrangements_for_exam_grading_and_assessment_in_2020.pdf and https://www.gov.uk/government/publications/awarding-qualifications-in-summer-2020#summer-symposium

Wrong question, wrong answer

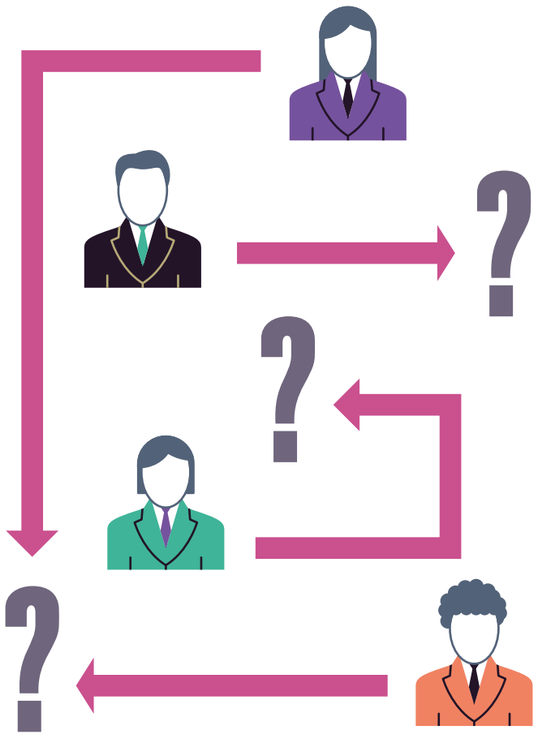
The solution was wrong because policy was trying to fix the wrong problem. We did not go wrong because the algorithm malfunctioned, but because human decision making went awry.

Cancelling exams made it impossible to know what grades people would have got. Suddenly we had lost the information we needed to decide who should go where in September. For policy makers and administrators, the question was, how do we find a way of enabling young people to progress? From this angle, the question quickly becomes how to fill the information gap left by exams, what information can be used instead, how to make sure the system doesn't break down.

Framing it this way assumes certain constraints: that the mechanism of allocating places stays the same and that the number of available places does not change. This is how the problem was considered in meeting rooms in Whitehall, Edinburgh, Cardiff and Belfast. How do we plug the gap in the machinery and work out which pupils go in which places? If that is the problem, moderated teacher grades are the correct answer. That is why all four administrations came to the same view. It is in this sense that moderated teacher assessments are 'the fairest thing to do', as I wrote at the time. They are the fairest solution to that problem.

> For policy makers and administrators [...] the question quickly becomes how to fill the information gap left by exams, what information can be used instead, how to make sure the system doesn't break down

> From the point of view of the individual citizen, the problem looks different. They see that the government has denied them the chance to demonstrate that they deserve a university place; it has put them at risk of unfairly losing out; it has put their future at risk.



From the point of view of the individual citizen, the problem looks different. They see that the government has denied them the chance to demonstrate that they deserve a university place; it has put them at risk of unfairly losing out; it has put their future at risk.

Let us imagine that the problem facing policy makers had instead been framed in these terms: How can government compensate young people for the fact that our policies mean they cannot produce the evidence they need to claim their place at university? Algorithmically moderated teacher grades do nothing to solve that problem. The algorithm does not compensate for or reduce their loss; it simply provides a defensible way to spread the pain.

The fact that there are, statistically, as many winners as losers makes no odds to the individual. One student's good fortune does not compensate another for their bad luck. More importantly, those who do benefit will be unaware of it, understandably thinking they deserved the results they got. The pain of disappointment falls on all students who were refused places, not just those incorrectly refused. It must do, since no-one can know the results of exams that never took place. Any student denied their hoped-for university place was left with the sense that the government's actions had robbed them of something that should have been theirs.

Statistically, from the point of view of someone operating a system of moderated grades, there were as many winners as losers. But for the individuals affected there were only losers.[10]

From this perspective, the argument that no-one knew exactly how many grades would change until they saw the results doesn't cut much ice. If it is unfair to refuse a university place on evidence that lacks the legitimacy and accuracy people were entitled to expect, it is no excuse to say we did not realise how inaccurate it would be. If anything it sounds even more cavalier.

A great deal of consideration was given to questions of compensation or redress for people who felt they had been given the wrong grade. There was an appeals process which could have fixed the 'obviously wrong' results. But this would not have helped the much larger number of students who felt, correctly or incorrectly, that they would have done better in an exam. For them, the answer was that they could sit exams in the Autumn.

> Statistically, from the point of view of someone operating a system of moderated grades, there were as many winners as losers. But for the individuals affected there were only losers.

This did, in theory, provide a means to correct a wrong result and was recognised as a crucial element in making the overall approach acceptable. In focus group research, the public were unequivocal in rejecting it. Most people regarded it as irrelevant to the fairness of summer awarding. Given the burden it places on the individual and the consequences for progression it was optimistic to imagine that it might have been seen as an acceptable way to correct a 'wrong' grade.

As long as the number of grades and the number of university places remained the same as before the pandemic, many students would be wrongly denied university places and many more would believe this had happened to them. The one policy that would compensate people for the cancellation of exams was to expand the number of university places. A significant increase would cater for those with a reasonable claim that they might have got a place if exams had taken place. It would acknowledge that, through no fault of their own, they were going to be unable to provide the evidence that is usually required. Allowing a much larger number of students to be admitted would limit the number who were wrongly excluded.[11] This option was, to my knowledge, never seriously considered. But by a painful, chaotic and unplanned route, it is where all four countries ended up.

The scrapping of moderated grades last summer was not the replacement of an unfair algorithm with a fair system of teacher grading. In terms of the distribution between groups there is little to choose between them. Abandoning moderation was the fairer thing to do because it lowered the threshold at which students were awarded a place at university. It was a way of acknowledging that, since it was government that had denied them the chance to prove they deserved a place at university, it was government that should make good the loss.

10 Social psychology has relevant insights here. Research into loss aversion shows that people put less value on unexpected gaining than they do on avoiding unexpected loses.

11 Further evidence that this was the primary issue comes from France and Ireland. Both were able to implement (albeit with challenge and controversy) systems of moderated teacher assessed grades. In France the moderation was done by local panels of adjudicators. In Ireland an adapted algorithm was used to apply a modest level of moderation. Both approaches allowed grades to inflate. This was accommodated by increasing the number of university places.

Learning the right lessons

Blaming an ‘algorithm’ suggests the problem was a feature of the maths, that some better form of algorithm would have fixed the problem, or that you can’t trust algorithms to do what you want. None of these conclusions would be true.

There are always legitimate arguments about how an algorithm affects distributions between groups and which is fairest. But it is implausible to suggest a different execution to achieve the same end would have won approval.

It is always right, as the Office for Statistical Regulation recommends, to ask whether an algorithm in any form is the right way to fix something. The ways that algorithms affect the accountability, legitimacy and fairness of decision making are vital considerations that will often count against their use as a solution. These were all important aspects of the events of 2020.[12]

But before any of these considerations are the questions, Is what this policy is trying to do reasonable? Is it fair to expect people to accept it? The failed attempt to replace cancelled exams with algorithmically moderated grades was first and foremost a colossal error of judgement about what people regard as acceptable and fair.

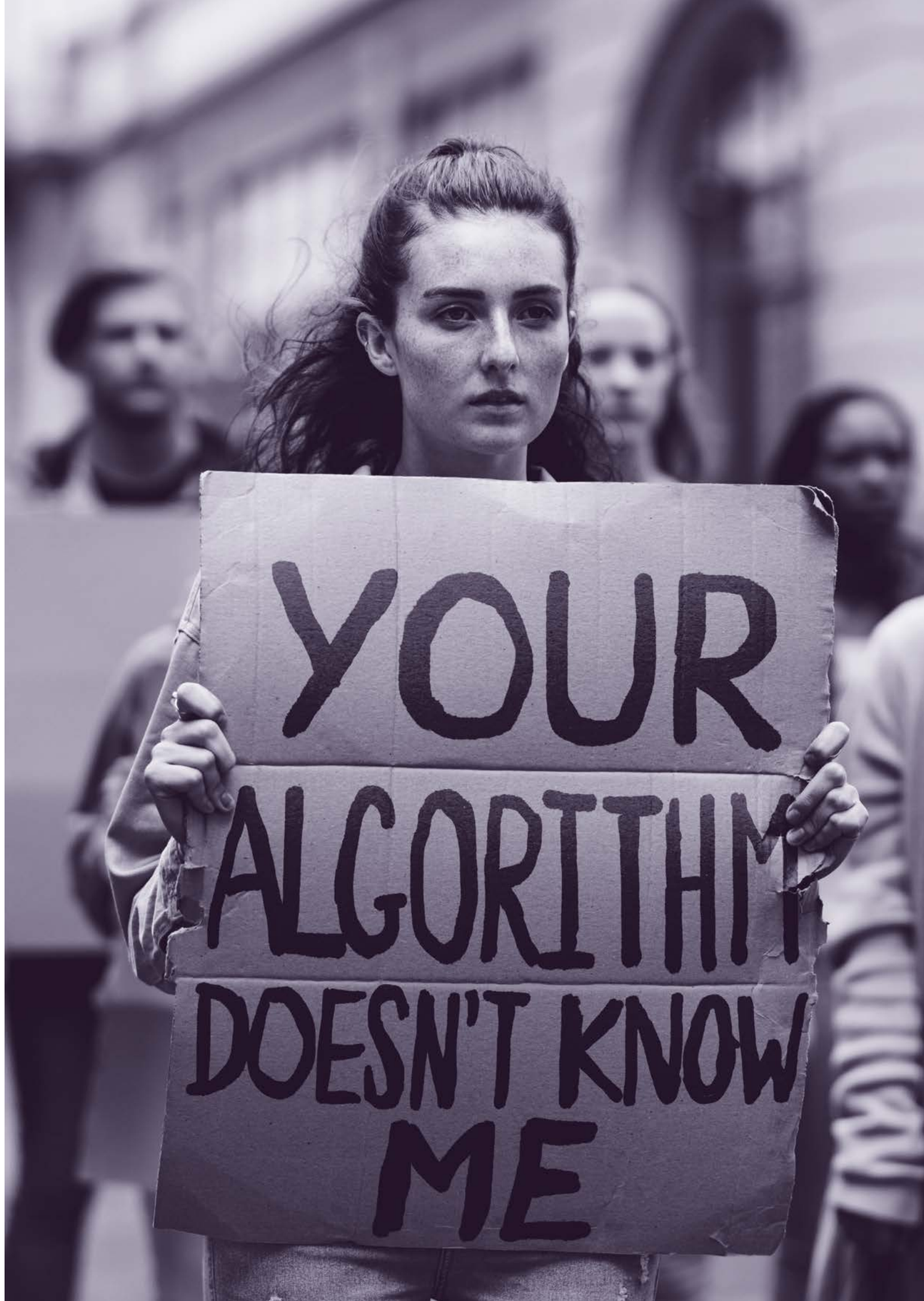
There is limited value in counterfactual speculation but I think it might have made some difference if, at the outset, there had been a clearer understanding that cancelling exams was to take something away from people they had a right to – a legitimate decision process. It might have helped if questions were posed more starkly in deciding the overall approach, such as, What is the remedy for the large number of people who are going to be wrongly denied their university places? The risks might have appeared more vivid if there had been a blunt acknowledgement that no remedy acceptable to the people affected was on offer.

If nothing else, it might have prompted ministers to ask, How many are we talking about? and, Is there really nothing else we can do? It might have opened a proactive discussion on whether it would be fairer to allow many more people to progress. It would also have enabled a discussion of how this could be achieved, and whether it could be done without issuing heavily inflated A-level grades.

> The events of 2020 have only reinforced that [...] people are not willing to accept their lives being affected by a decision-making process driven by predictive algorithms that imposes, or appears to impose, significant risks on them. We risk missing this very basic lesson if we comfort ourselves with the idea that the algorithm malfunctioned.

The question did arise in discussions regarding the level of ‘generosity’ in the approach to moderation. Ofqual’s decisions that gave students the ‘benefit of the doubt’ and allowed for modest inflation in grades was seen as an effort to make the process more palatable. But this is very different from a discussion about whether it is reasonable to keep the numbers progresing to university capped when no one can say with certainty who might legitimately deserve a place.

It was clear long before 2020 that public authorities should be cautious of adopting algorithmic decision systems unless they can demonstrate a substantial benefit over and above what went before.[13] The events of 2020 have only reinforced that lesson and emphasised that, even in the extreme circumstances of a pandemic, people are not willing to accept their lives being affected by a decision-making process driven by predictive algorithms that imposes, or appears to impose, significant risks on them. We risk missing this very basic lesson if we comfort ourselves with the idea that the algorithm malfunctioned.



12 A non-algorithmic solution to the problem of awarding the originally agreed number of university places would have been to award teacher assessed grades but allow universities to rescind their offers and make new offers in the light of those results. This would have given universities flexibility in deciding how many students to accept and enabled a more controlled increase. By putting the final decision in the hands of the universities it would quite likely have had greater legitimacy. But it still feels a something of a stretch to think this would have been acceptable. (This was, in effect, the Ofqual proposal to issue school leaving certificates. Leaving certificates would have meant that inflated grades did not conflict with the duty to maintain grade standards. It would have also meant universities were not obliged to meet offers and could have had more control over exactly they wished to admit within an overall increase in numbers.).

13 This applies primarily to systems that replace legitimate human judgement or allocate scarce resources. For purely transactional systems the concerns are less acute.

How algorithms are making decisions fairer

The debacle has given algorithms a reputation as a mechanism of injustice. This is unfortunate and undeserved. Every day, with little fanfare, algorithms and data-driven systems are used to make sense of exam results and address disadvantage. In these contexts, the algorithm is used to interpret exam results, not to replace them, helping employers and university admissions officers make fairer decisions. In Australia, the state of Victoria used an algorithm with public support to adjust exam results for lost learning.

Public confidence in exams was both damaged and bolstered by the events of the pandemic. The absence of exams reminded people of their value and this year, there has been an increase in people's confidence in exams.[14]

But at the same time, the events have highlighted concerns about examinations, leading to calls for reform. The issues raised are not new: fairness, accuracy, relevance and value.

> To many people, the accuracy of the algorithm was as objectionable as the inaccuracy. Putting into hard code the fact that the school you attend is a predictor of the grades you are likely to achieve does not seem to square with meritocracy.

Making fair comparisons

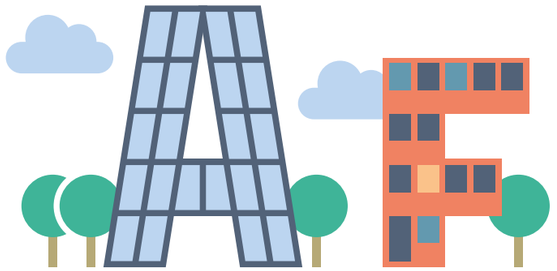
Qualifications and examinations were designed as instruments of meritocracy. They provide an objective playing field on which all compete equally, demonstrating their level of knowledge and skills as an indicator of their fitness for work or study. The way that qualifications have been handled in the last two years has exposed the extent to which this is not true.

People hated the algorithm because it limited the grades people could get on the basis of the school they attended. Watching as events unfolded in the UK, the government of the Republic of Ireland, which had been planning to use a similar approach, changed course at the last minute and decided not to use data about a school's past results as a factor in their algorithm.

The Irish government was taken to court by pupils from high-performing schools who were awarded grades lower than those they believed they would have got.[15] Changing the algorithm does not change the reality that the school you attend is a predictor of your grades.

To many people, the accuracy of the algorithm was as objectionable as the inaccuracy. Putting into hard code the fact that the school you attend is a predictor of the grades you are likely to achieve does not seem to square with meritocracy.[16]

This year, exams have been cancelled because it would be unfair when children have had such different levels of access to education, with lessons disrupted to varying degrees across the whole country. This approach has received wide support. But it prompts the question of why exams are fair in normal times, when pupils also experience very different levels of educational support, differences that are reflected in the exam results they achieve.[17]



14 Ofqual perceptions survey 2021. Available at: https://www.gov.uk/government/news/perceptions-of-qualifications-in-england-wave-19

15 A test case, brought by Freddy Sherry, a pupil at a fee-paying school in Dublin with a history of high grades, claimed that the government was wrong to remove school past performance from the algorithm. This claim was rejected in the high court. The judge ruled that the government was entitled to do this in order to maintain public confidence. https://www.irishlegal.com/article/high-court-student-fails-to-overturn-calculated-grades-from-the-2020-leaving-certificate. The simultaneous removal of the mechanism to prevent grade inflation meant protests were relatively muted.

16 The fact that the teacher-assessed grades built in the same expectation based on the teacher's knowledge of the grades that the school would likely have achieved is not generally regarded as objectionable.

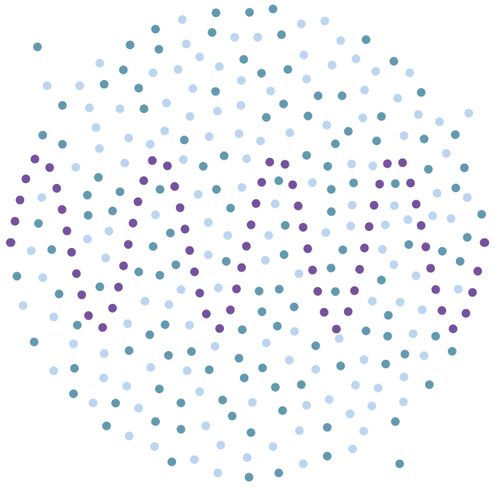
17 The most important response to these issues are to level-up access to high-quality education. Nothing in this paper is questioning that this is the overriding imperative. However, my topic is qualifications and there is no prospect that improvements in the education system will eradicate the problem that exam results can reflect how much help you had preparing.

Some qualifications, like driving tests, are used only to find out whether you can do something. It does not matter how much instruction you needed to pass the test, all that matters is whether you can drive. Others are used to help people decide how well you will cope with a university course or a job. This is not a hard distinction, but general qualifications are, in the main, in the second category. People care that you got an A in your history papers not because they need to know you can write a coherent account of the Tudors, but because they believe it may indicate what you are capable of. If your grades reflect your past advantage rather than your future potential, it may give them the wrong signal.

The people dealing with this issue are university admissions officers and employers. For the last two decades, universities have adopted a range of strategies to address this, including outreach support services and the use of contextual information.[18] Contextual data can be used to change the way grades are interpreted. For example, Cambridge University has used an algorithm [19] that scores an A grade between 0 and 1 depending on how often pupils at the candidate's school normally get A grades. This type of algorithm can then be used to adjust the thresholds required for an interview, or to make contextual offers where the grades required are lowered to reflect what exceptional performance looks like at your school. At the extreme, this approach can result in unconditional offers, where the university agrees to take the candidate regardless of their grades.[20]

Employers also face the problem that qualifications may say more about your social position than what you are capable of. One widespread practice is to delete information about qualifications, such as university degrees, from CVs on the grounds that a better degree from a more prestigious university may say more about the candidates' background than their abilities.

> Recruiters are using algorithms to sort out the signal from the noise in qualifications, to identify when grades indicate ability rather than advantage



Some employers are using methods of contextual recruitment similar to universities. RARE[21] is a service that allows grades from different candidates to be adjusted to take account of their advantage or disadvantage in terms of schooling and background. RARE is popular among elite recruiters such as leading law firms and consultancies, recruiters who have traditionally looked for candidates with very strong qualifications but found that this prevents them finding talent among more diverse and less advantaged groups.

For these firms, the need to diversify recruitment is not a reputational issue but an operational one. Failure to diversify means other firms hiring talent you missed; it means your firm will struggle to understand the diverse clients that it serves; its thinking will be constricted by unchallenged assumptions. The view that diversity is essential to business success is reflected in research showing an association between diversity and business success. These recruiters are using algorithms to sort out the signal from the noise in qualifications, to identify when grades indicate ability rather than advantage. This means recognising that the candidate who got the best mark in their borough, the candidate who stood out in their school, the candidate who completed their qualifications in the face of significant challenges offers greater potential than the candidate with perfect grades from a top school.

Box 1: Looking for rare talent

RARE was developed to help elite employers diversify recruitment and is used by a wide range of leading law firms, consultancies and professional services businesses. It uses information about candidates' grades alongside information about school performance and personal circumstances to calculate a 'performance index': an indication of how impressive the grades achieved are, given the level of educational support candidates have received.

Candidates provide information on their circumstances when growing up, such as whether they were working or whether they were a carer. As well as the performance index, the tool provides flags to identify particular issues, for example people who grew up in care or people from the lowest performing 10% of schools in the country. The tool helps employers gain a better sense of where to find the talent they are looking for. In particular, they can better identify exceptionally able people from difficult circumstances, for example candidates who were far ahead of their peers and their school in a particular subject, but whose grades on their own look unexceptional

18 See in particular the 'Schwartz report' 2004 on the difficulty of defining 'merit' and the ned to consider context. Available at: https://dera.ioe.ac.uk/5284/1/finalreport.pdf

19 This was being used in 2013 as described in https://www.rarerecruitment.co.uk/static/research/2013_Social_Mobility_in_Graduate_Recruitment.pdf. I do not know if this is currently what they do. It would be entirely reasonable if their current approach is not public information. People operating algorithms of this sort have obligations of accountability and transparency. But publishing details of exactly how their system works can be counterproductive as the damaging consequences of gaming are likely to outweigh any increase in accountability.

20 Unconditional offers can also be used as a marketing ruse to persuade candidates to accept university places when they would be better off doing something else, either going to a better university or doing something else entirely. In this paper, I am only looking at their use as a mechanism to encourage diversity. Offers of this sort assume the candidate is someone who could get good grades and, in that sense, endorse the view that grades matter. But they also imply that the university does not need to see the actual grades to confirm their assessment.

21 See: https://www.rarerecruitment.co.uk/

Do qualifications tell recruiters what they need to know?

Some employers have gone further. EY, the accountancy and consultancy business, announced some years ago that it was dropping all qualification requirements for its graduate recruitment stream and now relies primarily or exclusively on its own assessment process. Others are moving in the same direction.

These employers have started to ask whether qualifications are of any value at all in making recruitment decisions. They are turning instead to a wide range of alternative assessment mechanisms, many developed by companies in the recruitment industry. Some of these are conducted online; others are conducted at assessment centres. For some large graduate recruiters most applicants will be screened out using online automated tests, long before any human has glanced at their CV and without their qualifications ever having been considered.

These employers are not only interested in diversifying recruitment. They are looking for information about what candidates can do that they cannot get from qualifications. They may not be interested in whether someone can do trigonometry but whether they can read a graph; they are less concerned about a candidate's ability to structure an essay but would like to know that they can craft an email. Your A-level grades or 2.1 in English Literature may not tell them what they want to know.

It is important not to overstate this. Qualifications still form a core component of most recruitment systems, including into professional and executive roles. Even when employers ignore qualifications, the educational achievement that a qualification represents is often the most important factor in being recruited. The problem is sometimes nothing more than the limited information conveyed by a qualification grade. Good grades reflect natural aptitude, dedication and hard work, as well the education and support the candidate received.

EY has done a great deal of work to understand the relationship between qualifications and success at work. Their experience confirms that academic qualifications are relevant. Candidates who get over 300 UCAS points are significantly more likely to do well in their accountancy exams. But then so are candidates who get less than 300 points but do well on their numerical aptitude test.

Box 2: Graduate recruitment open to all

EY, like many large accountancy firms, used to have minimum academic requirements for anyone wanting to apply. Five years ago, these were dropped, in part because they acted as a barrier to diversity but also because EY found performing well in academic exams was only one predictor of whether a candidate would perform well at work.

35

Over 35 data points are collected to inform EY's hiring decisions

The first year saw a big change in intake. 18% of recruits would have been rejected under previous arrangements. After two years, a dip in performance in the accountancy exams led EY to re-evaluate the approach. Analysis of what would have predicted these outcomes showed that candidates with low UCAS points were unlikely to perform well. However, this was not true if they had achieved a good score on EY's numerical aptitude tests and EY now use both of these data points in identifying candidates. The approach has enabled EY to diversify recruitment and identify talent among candidates they would have previously rejected.

80%

Online tests and automated interviews screen out 80% of EY's applicants

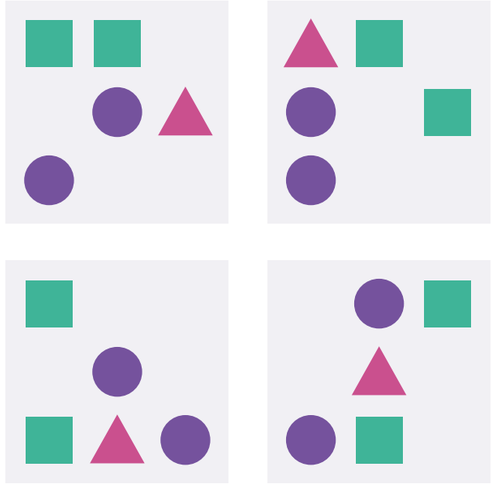
The full recruitment process includes online tests and automated interviews which screen out 80% of applicants. Those that progress attend an assessment centre where they again take a range of written tests, interviews and observed activities. Over 35 data points are collected to inform hiring decisions

How much does this matter?

There is much to lose if qualifications become less relevant to getting a job. This is not a criticism of what corporate recruiters are doing; they are motivated by a desire to increase the accuracy and diversity of recruitment and they are succeeding. However, in the long run it risks the opposite.

Regulated qualifications have characteristics that make them fairer than recruitment assessments. The knowledge and skills being assessed, and the mechanism by which they will be assessed are public. The candidate has certain rights, for example to see their exam paper after it is marked and to appeal. Public exams are part of an educational system in which assessments are linked to courses and to teaching. They are designed to support learning as well as selection; while recruiters' tools are focused only on selection. These qualities can also be found in many unregulated qualifications, for example in the IT industry, but they are often absent from recruitment assessment methods.

If employment decisions are made on the basis of less transparent or proprietary assessment mechanisms – mechanisms that do not connect as easily to the education provided to children – it risks eroding the value of qualifications and diminishing their currency. This would be a problem for everyone, for employers and government as well as candidates.



> Recruitment assessments may try to be less influenced by how candidates were educated; but if the candidates had not been educated they would not be employable

Employers rely on the education that a good qualification represents. Assessments used in recruitment are sometimes presented in the tradition of 'aptitude tests' that claim to be less affected by your education and better able to capture an underlying ability. However, it is never that straightforward. All tests of this sort are coachable – hence the many services offering to help for a fee. Numerical aptitude tests ask about things that are acquired in school, for example the ability to read a graph. If a firm is lucky enough to have lots of candidates score well on verbal reasoning tests, it will be because those candidates have been to schools where they were taught subjects such as history and biology. Recruitment assessments may try to be less influenced by how candidates were educated; but if the candidates had not been educated they would not be employable.

Another risk is cost. If a recruiter relies more on their own recruitment assessments, they then have to bear the cost of policing the system. Trying to ensure the ongoing validity of recruitment assessments is a constant battle against a thriving online community of people swapping tips on how to get through tests used by different firms, trading screenshots of online assessments and selling model answers. This is a major headache for recruiters.

Qualifications ought to be able to solve these problems for recruiters. If this is not happening, the onus is on those of us working in qualifications to address the issue.

Why is this happening?

Why is it that a firm like EY, which used to rely on qualifications as the starting point for selecting candidates, has moved so far in its approach to recruitment? Is it because the skills they need are changing? This is certainly an issue for a large number of people. The proportion of jobs for which so-called 'workplace skills' are essential is growing. These are behaviours such as resilience and team-working as well as more academic skills such as problem solving, critical thinking, literacy and numeracy. More and more people are going to need these skills as work changes.

But it does not explain the behaviour of firms like EY, or of recruiters hiring into law firms, consultancies and a wide range of executive jobs traditionally taken by graduates. They have always needed people who are numerate and literate on the one hand and, on the other, able to work with clients and colleagues. The skillset these firms value has not fundamentally altered.

The recognition that diversification is critical to business success has been a big shift. That is undoubtedly driving firms to think about how they can more accurately identify potential, not just achievement.

> The computerised systems that reject most candidates applying to professional firms will rapidly collect and process a wider and more detailed set of data about the candidate than they would get from their qualifications

The other change is digital technology. In the past, it made sense to recruit people on the basis of decent qualifications, a good interview and a reference, because that was about the amount of information the person responsible for recruitment could sensibly process.

Today, recruitment decisions are driven by more complex sets of data and computerised data processing. The systems that reject most candidates applying to professional firms will rapidly collect and process a wider and more detailed set of data about the candidate than they would get from their qualifications. Recruitment management platforms draw in scores from banks of tests, automated online interviews, observation of people working on complex tasks and interviews with psychologists, potential colleagues as well as the recruitment panel. This more granular information helps employers identify the people they want and mitigates the risk that qualifications steer them towards less able candidates with more advantaged backgrounds.[22]

This data-heavy approach continues once people are employed with a range of personnel management systems that hold detailed information about performance that can inform decisions about training.

There are many signs of the changes to recruitment caused by digital technology. The role of platforms such as LinkedIn is one. Another is the growth of digital credentials that a candidate can provide to potential employers. This is not simply more convenient than trying to share paper certificates. It can enable verification of information using block-chains and it can support a much wider range of information, such as micro-credentials. Micro-credentials provide information about very specific assessments and are popular with providers of online courses.

These are all developments that regulated qualifications need to adapt to. Where they do not, they risk being displaced in recruitment decisions by other forms of information and assessment.

22 All the information collected will quite likely reflect social advantage and may be better understood in context. However, used well additional information can enable fairer decisions.

Qualifications in a data-driven age

It is foolish to try to predict how the growth of AI and data-driven technologies will effect the way in which people are assessed and recruited. These changes raise a wide range of ethical issues and are engaging the attention of researchers and policymakers worldwide, including at the CDEI. It is for the qualifications industry to work out how best to adapt. Below, I have described four areas where government policy could support which would help create an environment that would support innovation and encourage the use of qualifications in fair recruitment.

1

Helping recruiters make sense of qualifications

It made sense to summarise a qualification in a single grade when the information was processed by a person reading a CV. It makes much less sense when the processing is being done by computers. A handful of letters and numbers is hardly an adequate summary of what a child has achieved over 14 or more years of education.

Digital certification and portable electronic education records allow for more detailed information about qualifications as well as relevant contextual information in a verified format that candidates could chose to share.

More granular information about qualifications is used today to a limited degree; for example, some law firms use information about candidates' rank within their year group on particular modules such as corporate law. When A-levels and GCSEs used a unified mark scheme, some universities used the scores to further refine their decisions. Provision of this information in standard verified digital forms might help recruiters give more weight to qualifications.

More detailed information would also help address the 'cliff-edge' that affects candidates who fall short of a grade boundary by just one or two marks, but who could not be said with any confidence to be significantly worse than the candidate who passed by one mark.[23]

Digital education records can make it easier for recruiters and admissions officers to understand the context around qualifications. They can support the provision of relevant validated information, perhaps candidates' work experience, health difficulties, or contribution to extracurricular activities. User-controlled digital education records would allow the results of non-qualification assessments, such as those conducted in assessment centres, to be recorded alongside qualifications. A system of user-controlled records would support the growing interest in micro-credentialling among innovators in qualifications.

Box 3: Why portable digital education records matter

Our careers are increasingly shaped by digital systems. Recruitment opportunities are communicated through online platforms, CVs are processed through AI text readers, even interviews can now be conducted entirely automatically. These techniques are hugely efficient and have the potential to increase social mobility and equality of opportunity. But the risks of unintended consequences are significant.

A key foundation in making these approaches accountable to and supportive of humans is the extent to which they are driven by data that exists independently of the system. People selecting candidates must have freedom (within legal constraints) to decide how they select from candidates. But there are certain characteristics of the way this is done that make it human friendly. Examples might include: it is clear to people how you can qualify for a position; qualifying depends on qualities that, in general, people are capable of achieving (i.e. learnt skills, not inherent attributes); these qualities can be acquired and assessed independently of any recruitment process; they are publicly understood and their definitions publicly contestable; there is educational support to help people achieve these things. The list could go on.

Qualifications are the mechanism by which we achieve that. It is not sufficient to say that the education a qualification represents is more important than the qualification itself. That is true, of course, in one sense. But if the qualification itself loses traction in the process by which our futures are shaped, the ability of humans to understand and shape how education supports flourishing lives will be compromised as data-driven technologies become ubiquitous.

The first step towards making qualifications fit for the future is to require that the information contained in a qualification is held on user controlled digital certificates. With that in place, we will be in a position to chart a course towards a future in which AI and technology work to support human flourishing.

23 There is no optimal level of grading. Dividing candidates into ten grades increases the likelihood that a candidate with an A is significantly better than one with a B. But it will exaggerate the difference between candidates at the boundary. Giving candidates a mark out of a hundred (in effect 100 grades) means it is less likely a candidate with a higher score is better than one with a lower score but it diminishes the impact of being wrongly graded. Large grades are useful shorthand and can come to represent a standard of work. More specific information allows greater understanding of the relative performance of two candidates. The great thing about digital records is you can have both relatively easily. Also, this does not require the application of a unified mark scheme. Raw marks and rank scores can be useful.

2

Flexibility, comparability and standards

A long-standing problem in qualifications is the tension between flexibility and standards. All qualifications encompass a range of different elements – different knowledge domains, a variety of skills etc. An algorithm determines how marks from different elements of assessment are combined into an overall grade, dealing with issues such as the extent to which doing well in one part can compensate for doing poorly in another.

The more flexible a qualification – the greater the number of ways you can get a particular grade – the harder it is to say with confidence that a grade in that qualification means something consistent.[24]

This issue arises with regard to modular qualifications. These can improve access and support lifelong learning but they make it hard to establish comparable standards. The idea that modular assessments or micro-credentials acquired over time could be combined to form the equivalent of a larger qualification taken at one time is problematic because it implies an equivalence between things that are very different.

This problem is most acute if candidates are compared primarily on the basis of a summary grade. The tension between these objectives is reduced if decisions are driven by richer data – granular information about how the overall grade was achieved or through the labelling of grades to indicate important differences. Digital certification of qualifications is one way to provide this information in a format that others could process.

3

Reliability and validity of grading higher order skills

A 'reliable' assessment will give the same grade to candidates of equal attainment (or will give a candidate the same grade if they take the test twice). Some things are relatively easy to assess reliably, for example knowing the names of capital cities. More complex notions, skills such as critical thinking creativity, are more subjective. Assessing them requires human judgement; judgement that will differ between individuals. This increases the likelihood that you might have been graded differently if someone else had marked your work.

Reliability can be improved by laying down very specific sets of criteria that define what it is to be 'creative' or 'analytical'. It can be achieved by using more constrained assessment such as multiple-choice questions rather than essay questions. But these result in overly reductive definitions of what it is to be analytical or creative.

Failure to fix this causes damages the disadvantaged most. The replacement of reliable grading of higher order skills with tick lists and market schemes that specify what counts as 'creativity' result in the worst sort of teaching to the test. It encourages schools to drill kids into learning pre-set phrases or following pre-planned essay formats that enable the student to give the impression of having the necessary skills but little true understanding. Inevitably, it is the children who are most disadvantaged, and hardest to teach, whose education is most degraded by these processes. Rather than improving education, these methods can create the false appearance of a narrowing attainment gap.

This dilemma has been around for as long as qualifications have existed. Recruitment assessment centres deal with it by using structured interviews or trained independent observers, the same approach used for Associated Board music exams. But that is not financially viable for schools. Replicating it with teacher assessment would lack sufficient reliability to carry much weight in recruitment decisions, and would risk introducing biases into grading.

Comparative judgment is an ingeniously powerful technique for dealing with problems of this sort, using digital technology to efficiently gather many different views about the relative strengths of different pieces of work. Ofqual has been conducting initial research into how it could inform the setting of grade boundaries. It is a technique for which there are many potential applications.

This approach is a radical change from traditional marking and its use in awarding qualifications would need careful thought. For example, it raises questions about how appeals against a grade would be handled. But the potential for it to allow for a more authentic grading of higher-order skills should be actively pursued.[25]

24 This also increases the risks that schools will identify the least demanding way to gain the qualification and teach this. This is discussed further in Section 4.

25 For more see: https://www.nomoremarking.com/

Box 4: The joy of comparative judgment

The curriculum for English Language GCSE requires that children are taught how to write in different styles including writing 'persuasively'. It is a very good idea to teach children what persuasive writing and persuasive techniques look like, not only to help them argue their own case, but to make them aware of how others may be manipulating them.

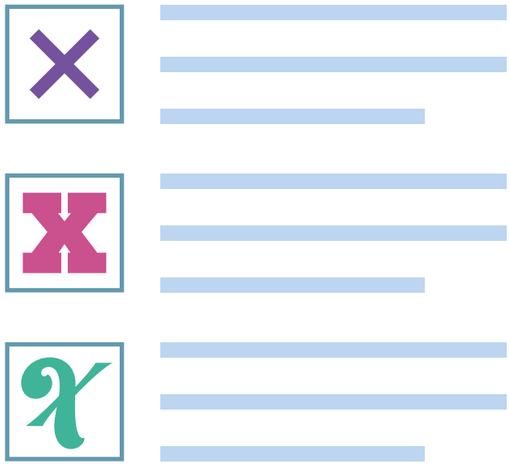
Under 'writing for impact' candidates are expected to show, amongst other things, that they can 'create emotional impact' and use language 'creatively and persuasively, including rhetorical devices' such as 'rhetorical questions' and 'antithesis'.

At one school this is taught by giving candidates a general purpose essay schema. The essay plan instructs them to use three sentences of the form: '"They say....; But we know.....', followed by three sentences each starting with: 'Imagine a world where.... ' and ending with the rhetorical question "Isn't that a world you would rather live in" . The students are trained to be able to generate a bit of text of this form in a range of contexts. This will ensure that their answer will include antithesis, a tricolon, a rhetorical question plus a colon and a semicolon. These all count for marks. So however weak the student's own creative input, there is a good chance they will get some marks. It is more debatable whether this counts as teaching people to communicate with impact, or whether persuasiveness can be reduced to notions such as using rhetorical questions. Can it really?

Marking guidelines help examiners know they are applying reasonably consistent standards. If they were simply told to give each answer a mark out of 10 on how persuasive they thought it was, it would leave too much room for interpretation. How would the marker know whether the papers they were marking represented work the normal range of abilities, or above or below average?

We are left choosing between random variability in marking or reductive systems of assessment. We tend to opt for the latter. This has a dramatic impact on teaching. If the way we assess 'creativity' or 'writing with impact' is reductive and inadequate, the way they are taught will become reductive and inadequate.

Comparative judgment holds out hope of solving this very fundamental problem in education. Instead of single marker reviewing an answer against a set of requirements, each marker is simply asked to compare two answers and say which they think is more 'persuasive'. Each answer is compared multiple times by multiple judges. The end result is a ranking of the answers with those higher up being answers which people felt, in the natural meaning of the word, were 'persuasive'. This ranking can then by matched to mark schemes of grades either through a similar process, comparing the work to work of an agreed mark standard, or through direct marking of a sample of texts by multiple markers.



Reliability [...] can be achieved by using more constrained assessment such as multiple-choice questions rather than essay questions. But these result in overly reductive definitions of what it is to be analytical or creative.

4

Choosing qualifications

Being smart about what to study and which qualifications to take gives those who know how the system works an advantage over others. The problem is most serious for those students who do less well academically where progression routes are more complex and harder to navigate.

This is another area where digital technology offers solutions. Simply presenting information on websites to help people make choices does little to level things up. But the use of digital targeting systems to contact young people can be a powerful aid to support other outreach mechanisms. These can engage students at little cost and show them how a particular course or qualification might help the person achieve something they had not thought was an option for them. Research by the CDEI on targeting found strong public support for the use of targeted information to make people aware of education, training and employment opportunities.[26]

> The use of digital targeting systems to contact young people can be a powerful aid to support other outreach mechanisms

Enabling technology

None of these ideas are new. But they could help to ensure qualifications do not lose the power they have had in the past as a force for social mobility. They can ensure qualifications play their full role in levelling-up. They can prepare our qualifications system for the increasing use of algorithmic systems in managing admissions and recruitment.

Complex information systems such as qualifications need to be continually improved by building on what works and adapting to changing needs. In contrast, the wholesale scrapping and replacing of systems is a more costly and uncertain route to improvement. The ideas set out above are all ways to allow the system to evolve and adapt.

The ideas set out above have two points in common. First, they do not reduce or degrade the quality or quantity of information that qualifications provide. People who are frustrated with aspects of qualifications are often drawn to proposals to scrap assessments or to adopt less rigorous forms of assessment – such as teacher assessment – in the hope that it will allow for a more holistic view of ability, or one that takes greater account of circumstances.

Such ideas are unlikely to work. If we reduce the reliability of information about candidates, it will only play to the advantage of the powerful. The introduction of less robust assessment tends to result in educational disadvantage being obscured, and alternative mechanisms being adopted to identify ability. The losers are those already most disadvantaged.

Second, the innovations described above are general purpose technologies. They are not things designed to help a government achieve a particular objective. They are mechanisms that can enable citizens, companies, employers, recruiters and anyone else with an interest achieve their objectives. That is why they have so much to offer.

26 See: https://www.gov.uk/government/publications/cdei-review-of-online-targeting

Governments vs citizens

At the start of this paper, I suggested that the key error in 2020 was misjudging what people would accept. I argued that this reflected a tendency of government and public administration to focus on the problem as is appears to them, with insufficient attention paid to the problem as experienced by the citizen. It reflected an attitude that data was there to help government solve its problems, not society.

This is admittedly a rather vague idea and warrants at least an attempt to describe it more completely. To do this, it is helpful to distinguish between two different issues. First there are those situations, such as the grading in 2020, where the problem is primarily about different perspectives. Government had no interest in promoting a policy that was unacceptable to people. It was simply a failure of imagination and policymaking that led to an outcome which everyone involved would have preferred to avoid.

A different situation issues arise when government sets objectives that conflict with the needs of individuals. Government objectives will often be designed to achieve a broad public benefit. But these may not coincide with the specific needs of individuals at a particular time.

Taking each of these situations in turn:

Understanding the perspective of citizens

One of the recommendations made in respect of 2020 is the need for **better public dialogue**, supported by appropriate methods of research and engagement.[27] This is an important part of the addressing the problem.

However, this approach has its limitations. A shared perspective and common assumptions contributed to the events of 2020. Research and consultation cannot be relied upon to dislodge ingrained predispositions in those conducting the research.

It was striking in the run up to 2020 that students in research groups could appreciate that without statistical moderation the results would not be fair. But this revealed nothing about the distress they felt the morning when so many discovered that they were the ones who were not going to get the grade they felt they could have achieved.

Another approach worth considering is a **more formalised process** to consider how changes in information systems affect the individual. When considering the risks and benefits of a system such as algorithmically moderated grades there are many questions to answer. Some are about the system overall: Will the number of losers/winners increase or decrease? Will certain groups be disadvantaged? These questions were looked at closely and were addressed.

Others focus more closely on the individual experience: Will there be people getting a worse result than they would have in the old system? Will they feel they have a legitimate grievance if this happens? Will it possible to show that the result is justified? How many people will believe this has happened to them? What would the cost be of allowing more people to progress to university? Many of these questions were considered and discussed. But self-evidently they were not given enough attention.

Lastly, the events of 2020 highlight the need to think explicitly about **legitimacy** when looking at the use of algorithms. Legitimacy is separate from technical considerations of accuracy, bias, explainability or recourse. Legitimacy is when people accept the authority of a decision-making process, errors, bugs, biases and all. Legitimacy is perhaps less a property of the decision-making process itself, and more something to be found in the attitudes and beliefs of the people affected by it.[28]

Human agency can play a big role in legitimacy – the degree to which your own actions determine the outcome.[29] This is an idea that arises in discussions of 'human centricity' and is central to an assessment of algorithmically informed decisions. But unlike issues such as bias which are relatively well-defined and assessed, notions of legitimacy can seem too nebulous to get much traction in policy discussions. That needs to change.

27 See: https://osr.statisticsauthority.gov.uk/publication/ensuring-statistical-models-command-public-confidence/

28 The point is that technical superiority does not, on its own, justify imposing an illegitimate system on people. It may be a good reason for trying to persuade people to change their views. And where a legitimate system clearly harms the rights of some people through bias or inaccuracy, there may be justification for overriding concerns. But legitimacy needs to be understood and factored into decision making.

29 A similar concern relates to the use of algorithms to inform bail decisions. There is evidence that such systems could improve the accuracy, consistency and fairness of bail decisions. But people may still prefer to be judged on their actions, their words, and their appearance in the court room.

Navigating conflicting aims between government and citizens

Qualifications serve a number of purposes. Governments use them as a lever to shape education and monitor performance. For students, their value is primarily in the ability to win them a job or a position for further study. For employers and admissions officers, the value is in enabling them to select talent.

These are all legitimate uses. But they do not wholly align. Government needs a system that can support accountability through consistent standard setting and comparability. Employers put more value on the ability to interpret qualifications and identify talent. For students, flexibility and second chances are highly prized, but less so if that flexibility reduces the value of their qualification.

Qualification design must make trade-offs between these objectives. No one objective is automatically paramount. A government might legitimately decide that the need to ensure young people get a high-quality education outweighs the needs of employers, who can use other mechanisms if they need more information to inform recruitment.

However, if the balance goes too far in this direction it becomes self-defeating. If the system appears to serve government to the detriment of other users, it undermines public support for the system and ultimately make it ineffective as a tool of policy. If people had to choose, they would put their efforts into getting a job, not satisfying a government target.

Data-driven technology can lessen the tension between these aims. The use of digital certification, electronic records, and granular reporting all have the potential to open up more space for qualifications to serve multiple ends simultaneously.

But for this to work, government needs to regard qualifications – and all similar national data systems – as primarily public utilities managed to support the different interests of users. This would have a number of practical consequences.

First, it would mean that wherever possible policy objectives should be achieved without resorting to mechanisms that distort information or restrict its use.

To illustrate this, modular GCSEs – GCSEs that could be taken in a number of small pieces over a longer time period – would be useful to adult learners and prisoners who would benefit from the greater flexibility. Government does not wish to see this option made available to schools because of the risk is that it would open-up a less demanding route to a GCSE. There is a risk that if it was available to schools it would appeal disproportionately to schools with more disadvantaged pupils resulting in a lower standard of education for those who most need the opposite.

The government requirement for linear qualifications to be taken in schools is currently implemented by requiring that GCSEs are linear. The tension between the interests of adult learners and the interests of government could be better managed if the policy requirement for linear qualifications were achieved by setting rules for the types of qualifications that should be taught in schools rather than restricting the form that a qualification can take.

Viewing information systems as public utilities has other consequences. The importance of maintaining standards is a clear public duty as no-one benefits from less reliable information. It implies that steps such as digital certification, which are useful primarily to students and recruiters, would be given greater priority than they are now. It implies that the knowledge contained in national data sets should not be monopolised by government but be available to civil society through appropriately managed research routes.

There is nothing very challenging here. These ideas are all currently reflected in various parts of the current administration's policies. Most have featured in government policy for the last two decades. That progress has not been faster is in part due to the complexity in implementing them.[30] But it is also slow because they get little priority, reflecting an attitude which underweights the value of qualifications as enabling infrastructure for citizens and overweights their role as a tool of government policy.

30 The progress that many in the private sector have made building the necessary infrastructure shows that the problems are often greatly exaggerated.

The errors of 2020 can be seen as a consequence of a particular mindset: one that is poor at recognising the citizens' perspective with regard to the impact of technology; and one that instinctively views data as a tool for the benefit of government. Fair and effective government needs AI and data-driven systems. These technologies have the potential to transform the way that the public sector operates for the better. However, this change will only be for the good if we get smarter about how to deploy them.

Government attitude towards data and data-driven systems is a fundamental source of mistrust. At the CDEI we have been surveying public attitudes to data usage. Most people (57%) understand that government bodies need to hold personal data in order to deliver services. But only 8% of people think they benefit from the government sharing data about them.[31]

That is remarkable. Government would not operate without sharing data about people. They do it to work out how to fund education, how to look after people's health, how much to tax people, where to build new roads, and how to keep people safe from pandemics. It is quite an achievement to have left most people with the impression that sharing data about them is not being done for their benefit.

There has been some optimistic commentary that Covid-19 may have made people more appreciative of how governments use data. I would not hope for too much. The survey I am quoting was done during the height of the pandemic. What happened with the Ofqual algorithm provides an important case study in why the public are sceptical.

Data-driven government is not, to most people, a pleasing idea. It brings to mind programmes that rob citizens of agency and create new hazards. It prompts thoughts of pointy-headed people making utilitarian calculations about averages, distributions and disbenefits. The problem is that, to a troubling degree, these impressions are correct.

This is a problem that was manageable when government digital systems for citizens were primarily transactional and where statistics were used to inform system level decisions – like budget allocations – rather than decisions about individuals. But we are now entering an era that will be dominated by the growth of algorithmic data-driven decision systems. In that context, a frame of mind that sees data as a tool for government rather than a public utility for the benefit of all will be toxic to good decision making and public trust.

Fair and effective government needs AI and data-driven systems. These technologies have the potential to transform the way that the public sector operates for the better. However, this change will only be for the good if we get smarter about how to deploy them.

8%

Only 8% of people think they benefit from the government sharing data about them

31 See: https://www.gov.uk/government/publications/local-government-use-of-data-during-the-pandemic

Appendices

Appendix 1: Explaining the Ofqual decision-making process

1. Why did Ofqual support the plan even though there was widespread recognition that it was likely to be rejected by the public?

Ofqual's primary statutory duty is maintaining qualification standards – making sure as far as possible that an A in A-level history or a BTEC in Art and Design means the same thing wherever and whenever you take it. Whether moderated or not, teacher-assessed grades would not meet that requirement and would be significantly less reliable than the current standards applied to GCSEs and A-levels. Ofqual put forward two possible ways forward that were consistent with its primary objective: hold exams in a socially-distanced environment or, alternatively, use 'non-qualification' leaving certificates to issue grades, while making clear they were not equivalent to A-level grades.

With some qualifications, such as functional skills, an alternative option was also used: adapting existing tests to the new environment, for example making them available online. However, this was not an option for qualifications with the scale and operational requirements of GCSEs and A-levels at a time when schools were closed.

It would have been possible for Ofqual to have stuck at that point and refused to allow A-level or GCSE grades to be issued without exams, on the grounds that they would lack validity. I don't think that would have helped the situation and it would have caused outrage. The view of the government was that neither approach recommended by Ofqual would command public confidence. I think that view is most likely right. Ofqual's remit allows it to act counter to its other duties in order to maintain public confidence. In our decision making, the position taken by the board of Ofqual was that the elected government has more legitimacy in deciding what will command public confidence than the regulator, and we would need exceptionally strong grounds to oppose them. For Ofqual to set itself in opposition to the government in an argument over what would command public confidence and about a policy which stakeholders supported would have been foolish.

Ofqual was able to propose that teacher assessed grades be used when the government recommended a course of action – grades based on mock exams – that was, in the view of Ofqual, an even less valid and fair mechanism for awarding university places than calculated grades or teacher assessed grades. It was on that basis alone that the board of Ofqual could decide that issuing teacher assessed grades was the right course of action.

2. Why did Ofqual not fix the 'obviously wrong' results in the moderated grades?

People are understandably mystified as to why Ofqual allowed some results to be awarded knowing that they would need to be changed on appeal. The reason for this was very strong legal advice that to make changes in advance of the award would quite likely result in the whole approach being rejected by the courts following one of the many judicial reviews that a number of law firms planned to request.

More than most other issues, this problem was a creation of the time-pressured circumstances, and with more time, problems of this sort could, possibly, have been resolved. But it may help to understand the particular issues that drove decision making. There were two moments. The first was early on in the process, when consideration was given to setting a maximum limit on how much the calculated grade could differ from the teacher grade, e.g. one or two grades. As only 2% of grades moved by more than one grade it would not have made that much of a difference to the overall results but would have prevented some of the 'inexplicable' changes.

This approach was rejected on principle, because it was not consistent with using the most reliable evidence available. It would have been an arbitrary rule that, in principle, would do more to increase errors than correct them. More pragmatic voices felt that this was an acceptable price to pay, if it meant people were more likely to accept the whole arrangement. To make it legally sound Ofqual would have needed to include in the guiding principles, one allowing for considerations of likely public acceptance. It is unclear how that would square with the other principles.

Once the results were calculated and it became clear that there were a fair number that 'looked wrong', work began on trying to identify how they might be corrected. A number of different definitions and categories were devised on the basis of particular instances – for example, students who were top of their school rankings with an unusually high grade for the school, where the school had not given most students unusually high grades but where the candidate's grade was none the less being moderated down as unlikely to have occurred. These definitions looked plausible but closer investigation led to two conclusions: first that there was no way of defining a rule based on observations of 'obviously wrong' results that could be shown to be an improvement overall. Any such rule applied to the algorithm would create arbitrary cut-offs and unintended effects. Subsequent work by Ofqual has shown the impossibility of coming up with a rule that could be demonstrated as an improvement without recourse to further information.

The second conclusion was that any arbitrary decisions that could not be shown to be implementing the principles consulted on, would make it difficult to defend the system in the face of planned judicial reviews. This led to the conclusion that changing individual results on appeal was potentially defensible (although even here there were concerns) while changing results in advance carried a high risk that the entire approach would be thrown out by the courts. The decision at the time was that Ofqual should act in a way that was defensible legally, even if it increased the risks of public rejection.

While the rationale at the time was clear, it was also evident that there was something amiss with a situation that prevented us as an organisation from fixing things everyone felt ought to be fixed. It might have been possible to establish in consultation authority to adjust the algorithm to fix problems that Ofqual felt, in its expert opinion, were likely to improve the reliability of results (but lawyers may differ on this point).

Outside of a public body working in a closely scrutinised environment, tweaking an algorithm on the basis of a human judgement that 'that can't be right' would be thought reasonable or desirable. Establishing clear legal processes for Ofqual and similar public authorities to do this – without diminishing accountability for those decisions – would have been helpful.

I would like to thank the many people who commented on early drafts of this paper and in particular: Michelle Meadows, Mike Cresswell, Richard Sargeant, Neil Carberry, Raph Makades, James Gordanifar, Sally Collier and Adrian Weller.

Designed by williamjoseph.co.uk

This paper is part of an occasional series for CPP of expert reflections on key questions of policy. Alongside it, CPP are publishing an analysis of inequalities in educational attainment with recommendations for ways in which our assessment system can change to better meet the needs of all young people, communities, employers and the wider economy.

All errors and omissions in the report are the responsibility of the author.

About the Centre for Progressive Policy

The Centre for Progressive Policy is a think tank committed to making inclusive economic growth a reality. By working with national and local partners, our aim is to devise effective, pragmatic policy solutions to drive productivity and shared prosperity in the UK.

Inclusive growth is one of the most urgent questions facing advanced economies where stagnant real wages are squeezing living standards and wealth is increasingly concentrated. CPP believes that a new approach to growth is needed, harnessing the best of central and local government to shape the national economic environment and build on the assets and opportunities of place. The Centre for Progressive Policy is funded by Lord Sainsbury and host of the Inclusive Growth Network.

Centre for Progressive Policy
27 Great Peter Street
London SW1P 3LN

+44 (0)20 7070 3360
www.progressive-policy.net