Office for Statistics Regulation

Ensuring statistical models command public confidence

Learning lessons from the approach to developing models for awarding grades in the UK in 2020

March 2021

Contents

Foreword

When students received their grades in August 2020 following the cancellation of summer exams in the UK, there was widespread confusion for students and public concern in all four countries as to the robustness of the grades that were awarded. Much of the criticism was focused on the role that statistical models or algorithms had played in the awards process.

Whilst the specific approaches adopted to award grades differed in the four jurisdictions, all dropped their planned approach and instead awarded grades based on teacher assessment of likely grades, where these were higher than the calculated grades. None of the planned approaches to awarding grades were able to command public confidence.

This report explores the approach to awarding grades in each of the four countries in order to identify wider lessons for government and other public bodies. In doing so, we recognise the constraints and unique challenges of this task. Context is crucial, and I want to emphasise four key points:

There can be an over-confidence in what statistical models can achieve: Statistical models have limitations. They are based on a number of assumptions and the data that are available. The results from statistical models are subject to variability. In the grade awarding context, the models were expected to predict a single grade for each individual on each course within constraints around maintaining standards and not disadvantaging any groups.

The task of awarding exam grades in the summer of 2020 was extremely difficult: It is important not to underestimate the scale of the task facing the qualification regulators. They are small bodies and faced a shortage of both resources and time. Moreover, they did not have perfect information with which to work – for example, the practice of ranking students within a cohort was not one for which there was any historical evidence or experience. Finally, unlike most other uses of statistical models, they had to release their output on a single day to an entire cohort and for use in processes such as university admissions.

There were important differences in approach between the four parts of the UK: There were similarities and differences in the models used in each country to calculate grades. These arose from differences in the design of qualifications, the data that was available and the decisions taken by the qualification regulators and awarding bodies. The biggest difference in approach was between Scotland and the rest of the UK countries, as the model in Scotland made greater use of teacher-estimated grades. Whilst more similar in approach, there were also differences in England, Wales and Northern Ireland, for example around the use of teacher rankings; the extent of quality review; and the use of prior performance of this cohort of students. The extent of differences between the four countries means that there is no single conclusion that can be drawn as to the best approach. Given this, this report does not attempt to evaluate or rank each approach.

It would be a mistake to think that the grade award decisions were made solely by a technical algorithm: Grades were also determined by teacher judgements, whether in the form of rankings and/or centre assessed grades. There were many other aspects to the awards process, including public engagement, and appeals, that were not the product of an algorithm. In addition, the use of statistical models to support the setting and maintenance of standards at the cohort level is a common feature of awarding grades in a normal year, but what was required of the standardisation approach was very different in 2020.

What this review has highlighted is that **achieving public confidence is key to the success of any model.** The development and implementation of statistical models in public services can never be just a technical exercise in coding. Technical features of a model are important, but alone they cannot secure public confidence. Instead, it is essential to have a rounded approach which builds on clear policy objectives, is appropriately governed, embeds transparency and quality in the process, and which considers ethical issues and public acceptability from the outset.

The distinction between technical judgements and broader public confidence is important in the context of the exams experience. In a normal year, people may not get the grades they expect, perhaps because they do not perform well on a particular day. And in a normal year, there is extensive expert moderation and standardisation of grades, drawing on statistical analysis. It is a well-established process, which may not be perfect, but which broadly speaking is accepted by the public as being an authoritative assessment of a student's performance.

In 2020, new approaches were developed, using new statistical models. These approaches had to command public confidence in the same way as a normal year. Securing this confidence depended not just on the design of the models themselves in technical terms. The regulators also faced a broader public scepticism about whether a heavily automated approach could be as authoritative as the more familiar system.

This was perhaps the greatest challenge facing the four regulatory bodies. They wanted the 2020 grades to command respect, and for people to regard them as comparable to grades awarded in 'normal' years. Yet they also were undertaking a novel approach, which, like any statistical approach, had limitations. They seemed to us to be wary of spelling out the limitations too strongly before the grades were awarded, as they felt this might undermine confidence. We understand this caution. However, given that the 2020 awarding process was in the public eye to a greater extent than a normal year, we consider that they should have placed greater weight on explaining the limitations of the approach.

In developing our lessons for government and other public bodies, we have sought not to bring too much hindsight to our analysis and findings. The approaches adopted by the regulators had many strengths. In some areas, though, we consider they could have made different choices, for example, around quality assurance and use of external expertise.

But the fact that the differing approaches led to the same overall outcome in the four countries implies to us that there were inherent challenges in the task; and these challenges meant that it would have been very difficult to deliver exam grades in a way that commanded complete public confidence in the summer of 2020.

Ed Humpherson

Director General, Office for Statistics Regulation

Executive Summary

Purpose of this report

In March 2020 the ministers with responsibility for education in England, Scotland, Wales and Northern Ireland announced the closure of schools as part of the UK's response to the coronavirus outbreak. Further government announcements then confirmed that public examinations in summer 2020 would not take place.

The four UK qualification regulators – Ofqual (England), Scottish Qualifications Authority (Scotland), Qualifications Wales (Wales) and the Council for the Curriculum, Examinations & Assessment (Northern Ireland) – were directed by their respective governments to oversee the development of an approach to awarding grades in the absence of exams. While the approaches adopted differed, all approaches involved statistical algorithms.

When grades were released in August 2020, there was widespread public dissatisfaction centred on how the grades had been calculated and the impact on students' lives. The grades in all four countries were re-issued based on the grades that schools and colleges had originally submitted as part of the process for calculating grades.

The public acceptability of algorithms and statistical models had not been such a prominent issue for so many people before, despite the rise in their use. As the regulator of official statistics in the UK, it is our role to uphold public confidence in statistics.

Statistical models and algorithms used by government and other public bodies are an increasingly prevalent part of contemporary life. As technology and the availability of data increase, there are significant benefits from using these types of models in the public sector.

We are concerned that public bodies will be less willing to use statistical models to support decisions in the future for fear of a public acceptability backlash, potentially hindering innovation and development of statistics and reducing the public good they can deliver. This is illustrated by the emphasis placed on not using algorithms during discussions of how grades will be awarded in 2021 following the cancellation of exams this year. For example, the Secretary of State for Education, when outlining the approach to awarding grades in January 2021, stated that "This year, we will put our trust in teachers rather than algorithms."[1]

It is important therefore that lessons are learned for government and other public bodies who may wish to use statistical models to support decisions. This review identifies lessons for model development to support public confidence in statistical models and algorithms in the future.

[1] The Secretary of State for Education, Covid-19: Educational Settings Volume 686: debated on Wednesday 6 January 2021, Hansard

The broader context: Models and algorithms

Throughout this report we have used the terms statistical model and algorithm when describing the various aspects of the models used to deliver grades. It should be noted, however, that terms such as statistical model, statistical algorithm, data-driven algorithms, machine learning, predictive analytics, automated decision making and artificial intelligence (AI), are frequently used interchangeably, often with different terms being used to describe the same process.

We consider that the findings of this review apply to all these data-driven approaches to supporting decisions in the public sector whatever the context.

Our approach: Lessons on building public confidence

This review centres on the importance of public confidence in the use of statistical models and algorithms and looks in detail at what it takes to achieve public confidence. The primary audiences for this review are public sector organisations with an interest in the use of models to support the delivery of public policy, both in the field of education and more broadly. This includes statisticians and analysts; regulators; and policy makers who commission statistical models to support decisions.

In conducting our review, we have adopted the following principles.

- Our purpose is not to pass definitive judgments on whether any of the qualification regulators performed well or badly. Instead, we use the experiences in the four countries to explore the broader issues around public confidence in models.
- The examples outlined in this report are included for the purposes of identifying the wider lessons for other public bodies looking to develop or work with statistical models and algorithms. These examples are therefore not an exhaustive description of all that was done in each country.
- In considering these case studies, we have drawn on the principles of the Code of Practice for Statistics. While not written explicitly to govern the use of statistical algorithms, the Code principles have underpinned how we gathered and evaluated evidence, namely:

  Trustworthiness: the organisational context in which the model development took place, especially looking at transparency and openness

  Quality: appropriate data and methods, and comprehensive quality assurance

  Value: the extent to which the models served the public good.

- We considered the end-to-end processes, from receiving the direction from Ministers to awarding of grades and the planned appeals processes, rather than just the technical development of the algorithms themselves.
- We have drawn on evidence from several sources. This included meeting with the qualification regulators and desk research of publicly available documents.

- We have undertaken this review using our regulatory framework, the Code of Practice for Statistics. It is outside our remit to form judgments on compliance or otherwise with other legal frameworks.

We have also reviewed the guidance and support that is available to organisations developing statistical models and algorithms to identify whether it is sufficient, relevant and accessible and whether the available guidance and policies are coherent. Independent reviews of the grade awarding process have been commissioned by the Scottish Government, Welsh Government and Department of Education in Northern Ireland. Whilst there are some overlaps in scope with our review, there are also key differences – most notably, the reviews sought to review the approach to awarding grades in order to make recommendations for the approach to exams in 2021. Our review goes wider: it seeks to draw lessons from the approaches in all four countries to ensure that statistical models, whatever they are designed to calculate, command public confidence in the future.

Findings

The approaches to awarding grades were regulated by four bodies:

- In England, Office of Qualifications and Examinations Regulation (Ofqual)
- In Scotland, Scottish Qualifications Authority (SQA)
- In Wales, Qualifications Wales
- In Northern Ireland, Council for the Curriculum, Examinations & Assessment (CCEA).

Although the specific approaches differed in the four countries, the overall concepts were similar, in that they involved the awarding of grades based on a mix of teacher predicted grade, rankings of students within a subject, and prior attainment of the 2020 students and/ or previous cohorts at the same centre (i.e. school or college).

It was always going to be extremely difficult for a model-based approach to grades to command public confidence

The task of awarding grades in the absence of examinations was very difficult. There were numerous challenges that the qualification regulators and awarding organisations had to overcome. These included, but are not limited to:

- The novelty of the approach, which meant that it was not possible to learn over multiple iterations and that best practice did not already exist.
- The constraints placed on the models by the need to maintain standards and not disadvantage any groups.
- The variability in exams results in a normal year due to a range of factors other than student ability as measured by prior attainment.
- Tight timescales for the development and deployment of the model.

- Decisions about young people’s lives being made on the day the grades were released.
- Limited data on which to develop and test the model.
- The challenges of developing the models while all parts of the UK were in a lockdown.
- Teacher estimated grades varied significantly from historic attainment for some schools or colleges.

These challenges meant that it was always going to be difficult for a statistical algorithm to command public confidence.

Whilst we understand the unique and challenging context in which the models were developed, we also recognise that the grade awarding process in summer 2020 had a fundamental impact on young people’s lives.

Public confidence was influenced by a number of factors

Against the background of an inherently challenging task, the way the statistical models were designed and communicated was crucial. This demonstrates that the implementation of models is not simply a question of technical design. It is also about the overall organisational approach, including factors like equality, public communication and quality assurance.

Many of the decisions made supported public confidence, while in some areas different choices could have been made. In our view, the key factors that influenced public confidence were:

The teams in all of the qualification regulators and awarding **organisations acted with honesty and integrity.** All were trying to develop models that would provide students with the most accurate grade and enable them to progress through the education system. This is a vital foundation for public confidence.

Confidence in statistical models in this context - whilst we recognise the unique time and resource constraints in this case, a high level of confidence was placed in the ability of statistical models to predict a single grade for each individual on each course whilst also maintaining national standards and not disadvantaging any groups. In our view the limitations of statistical models, and uncertainty in the results of them, were not fully communicated. More public discussion of these limitations and the mechanisms being used to overcome them, such as the appeals process, may have helped to support public confidence in the results.

Transparency of the model and its limitations – whilst the qualification regulators undertook activities to communicate information about the models to those affected by them and published technical documentation on results day, full details around the methodology to be used were not published in advance. This was due a variety of reasons, including short timescales for model development, a desire not to cause anxiety amongst students and concerns of the impact on the centre assessed grades had the information been released sooner. The need to communicate about the model, whilst also developing it, inevitably made transparency difficult.

Use of external technical challenge in decisions about the models - the qualification regulators drew on expertise within the qualifications and education context and extensive analysis was carried out in order to make decisions about the key concepts in the models. Despite this, there was, in our view, limited professional statistical consensus on the proposed method. The methods were not exposed to the widest possible audience of analytical and subject matter experts, though we acknowledge that time constraints were a limiting factor in this case. A greater range of technical challenge may have supported greater consensus around the models.

Understanding the impact of historical patterns of performance in the underlying data on results – in all four countries the previous history of grades at the centre was a major input to calculating the grades that the students of 2020 received for at least some of their qualifications. The previous history of grades would have included patterns of attainment that are known to differ between groups. There was limited public discussion ahead of the release of results about the likely historical patterns in the underlying data and how they might impact on the results from the model. All the regulators carried out a variety of equality impact analyses on the calculated grades for potentially disadvantaged categories of students at an aggregate level. These analyses were based on the premise that attainment gaps should not widen, and their analyses showed that gaps did not in fact widen. Despite this analytical assurance, 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.

Quality Assurance – in the exam case, there were clear examples of good quality assurance of both input and output data. For input data, centres were provided with detailed guidance on the data they should supply. For output data, the regulators undertook a wide range of analysis, largely at an aggregate level. There was limited human review of outputs of the models at an individual level prior to results day. Instead, the appeal process was expected to address any issues. There was media focus on cases where a student's grade was significantly different from the teacher prediction. In our view, these concerns were predictable and, whilst we recognise the constraints in this scenario, such cases should be explored as part of quality assurance.

Public engagement – all the qualification regulators undertook a wide range of public engagement activities, particularly at the outset. They deployed their experience in communicating with the public about exams and used a range of communication tools including formal consultations and video explainers, and the volume of public engagement activity was significant. Where acceptability testing was carried out, however, the focus was primarily on testing the process of calculating grades, and not on the impact on individuals. This, and the limited testing in some countries, may have led to the regulators not fully appreciating the risk that there would be public concern about the awarding of calculated grades.

Broader understanding of the exams system: in a normal year, individuals may not get the results they expect. For example, they may perform less well in an exam than anticipated. Statistical evidence and expert judgments support the setting of grade boundaries in a normal year. These may not be well understood in general but, as well-established processes they are able to command public confidence. As a result, when the unfamiliar 2020 approach was presented publicly, people may have assumed that an entirely new, machine-led approach was being introduced, and this may have raised their concerns. This issue of broader understanding would have been very hard for the regulators to address in the time available.

Overall, what is striking is that, while the approaches and models in the four countries had similarities and differences, all four failed to command public confidence. This demonstrates that there are key lessons to be learned for government and public bodies looking to develop statistical models to support decisions. These lessons apply to those that develop statistical models, policy makers who commission statistical models and the centre of government.

Lessons for those developing statistical models

Our review found that achieving public confidence is not just about delivering the key technical aspects of a model or the quality of the communication strategy. Rather, it arises through considering public confidence as part of an end-to-end process, from deciding to use a statistical model through to deploying it.

We have identified that public confidence in statistical models is supported by the following three principles:

- **Be open and trustworthy** – ensuring transparency about the aims of the model and the model itself (including limitations), being open to and acting on feedback and ensuring the use of the model is ethical and legal.
- **Be rigorous and ensure quality throughout** – establishing clear governance and accountability, involving the full range of subject matter and technical experts when developing the model and ensuring the data and outputs of the model are fully quality assured.
- **Meet the need and provide public value** – engaging with commissioners of the model throughout, fully considering whether a model is the right approach, testing acceptability of the model with all affected groups and being clear on the timing and grounds for appeal against decisions supported by the model.

Specific learning points, which are of relevance to all those using data-driven approaches to support decisions in the public sector underpin each principle. These are detailed in Part 3 of this report.

Lessons for policy makers who commission statistical models

We have identified lessons for ensuring public confidence for commissioners of statistical models from the perspective of supporting those developing them.

- **A statistical model might not always be the best approach to meet your need.** Commissioners of statistical models and algorithms should be clear what the model aims to achieve and whether the final model meets the intended use, including whether, even if they are “right”, they are publicly acceptable. They should ensure that they understand the likely strengths and limitations of the approach, take on board expert advice and be open to alternative approaches to meeting the need.

- **Statistical models used to support decisions are more than just automated processes.** They are built on a set of assumptions and the data that are available to test them. Commissioners of models should ensure that they understand these assumptions and provide advice on acceptability of the assumptions and key decisions made in model development.
- **The development of a statistical model should be regarded as more than just a technical exercise.** Commissioners of statistical models and algorithms should work with those developing the model throughout the end to end process to ensure that the process is open, rigorous and meets the intended need. This should include building in regular review points to assess whether the model will meet the policy objective.

Lessons for centre of Government

For statistical models used to support decisions in the public sector to command public confidence, the public bodies developing them need guidance and support to be available, accessible and coherent.

The deployment of models to support decisions on services is a multi-disciplinary endeavour. It cuts across several functions of Government, including the Analysis function (headed by the National Statistician) and the Digital and data function, led by the new Central Digital and Data Office, as well as others including operational delivery and finance. As a result, there is a need for central leadership to ensure consistency of approach.

The Analysis Function aims to improve the analytical capability of the Civil Service and enable policy makers to easily access advice, analysis, research and evidence, using consistent, professional standards. In an environment of increasing use of models, there is an opportunity for the function to demonstrate the role that analysis standards and professional expertise can play in ensuring these models are developed and used appropriately.

Our review has found that there is a fast-emerging community that can provide support and guidance in statistical models, algorithms, AI and machine learning. However, it is not always clear what is relevant and where public bodies can turn for support - the landscape is confusing, particularly for those new to model development and implementation. Although there is an emerging body of practice, there is only limited guidance and practical case studies on public acceptability and transparency of models. More needs to be done to ensure there is sufficient access for public bodies to available, accessible and coherent guidance on developing statistical models

Professional oversight support should be available to provide support to public bodies developing statistical models. This should include a clear place to go for technical expertise and ethics expertise.

Our recommendations

These recommendations focus on the actions that organisations in the centre of Government should take. Those taking forward these recommendations should do so in collaboration with the administrations in Scotland, Wales and Northern Ireland, which have their own centres of expertise in analysis, digital and data activities.

Recommendation 1: The Heads of the Analysis Function and the Digital Function should come together and ensure that they provide consistent, joined-up leadership on the use of models.

Recommendation 2: The cross-government Analysis and Digital functions, supported by the Centre for Data Ethics and Innovation should work together, and in collaboration with others, to create a comprehensive directory of guidance for Government bodies that are deploying these tools.

Recommendation 3: The Analysis Function, Digital Functions and the Centre for Data Ethics and Innovation should develop guidance, in collaboration with others, that supports public bodies that wish to test the public acceptability of their use of models.

Recommendation 4: In line with the Analysis Function's Aqua Book, in any situation where a model is used, accountability should be clear. In particular, the roles of commissioner (typically a Minister) and model developer (typically a multi-disciplinary team of officials) should be clear, and communications between them should also be clear.

Recommendation 5: Any Government body that is developing advanced statistical models with high public value should consult the National Statistician for advice and guidance. Within the Office for National Statistics there are technical and ethical experts that can support public bodies developing statistical models. This includes the Data Science Campus, Methodology Advisory Service, National Statistician's Data Ethics Committee and The Centre for Applied Data Ethics.

We will produce our own guidance in 2021 which sets out in more detail how statistical models should meet the Code of Practice for Statistics. In addition, we will clarify our regulatory role when statistical models and algorithms are used by public bodies.

Conclusion

The grade awarding process in 2020 was a high-profile example of public bodies using statistical models to make decisions.

In our view, the teams within the qualification regulators and awarding organisations worked with integrity to try to develop the best method in the time available to them. In each country there were aspects of the model development that were done well, and aspects where a different choice may have led to a different outcome. However, none of the models were able to command public confidence and there was widespread public dissatisfaction of how the grades had been calculated and the impact on students' lives.

Our review has identified lessons to ensure that statistical models, whatever they are designed to calculate, can command public confidence in the future. The findings of this review apply to all public bodies using data-driven approaches to support decisions, whatever the context.

Our main conclusion is that achieving public confidence in statistical models is not just about the technical design of the model – taking the right decisions and actions with regards transparency, communication and understanding public acceptability throughout the end to end process is just as important.

We also conclude that guidance and support for public bodies developing models should be improved. Government has a central role to play in ensuring that models developed by public bodies command public confidence. This includes directing the development of guidance and support, ensuring that the rights of individuals are fully recognised and that accountabilities are clear.

Introduction

The Office for Statistics Regulation aims to improve public confidence in the use of statistical models after the reduction in trust caused by the models developed to award grades in 2020.

Our vision is simple: **Statistics should serve the public good.** As regulators, it is our role to uphold public confidence in statistics by addressing harms and championing high standards to make sure that statistics serve the public good. These standards are embodied in the principles and practices of the Code of Practice for Statistics within the pillars of Trustworthiness, Quality and Value.

Background

During the coronavirus pandemic many difficult decisions have had to be taken. Everyone has had to balance their desire to continue with life as normal with the health of themselves and others. At the national level the decision was taken to close schools and workplaces, only allow essential shops to open and ask everyone to stay home as much as possible to minimise the spread of Covid19 and loss of life. As part of this, the decision was taken to cancel public examinations in all four countries of the UK and to use statistical models to award grades in 2020.

Prior to 'results day' there had been growing concern about the statistical models and calculated results, particularly around the impact on students from disadvantaged groups. Once the results were published headlines abounded about how statistics had 'ruined my life' and phrases such as 'mutant algorithm' were used. Students protested about the grades they had been awarded. The Governments and qualifications regulators in all four countries decided to re-issue the grades based on the grades that schools and colleges had originally submitted to awarding organisations as part of the process for calculating grades.

It was clear that public confidence in statistics had been damaged by the use of these statistical models and a review of what had happened was within our remit.

Scope of the report

This report outlines the findings of our review into the approach taken to developing the statistical models designed for awarding 2020 exam results. We have identified key learning for public bodies considering the use of statistical models to support decisions, those commissioning such models and the centre of government. We have focused on what lessons can be learned to ensure statistical models designed to support decision-making command public confidence.

In conducting our review, we adopted the following principles.

- Our purpose is not to pass definitive judgments on whether any of the qualification regulators performed well or badly. Instead, we use the experiences in the four countries to explore the broader issues around public confidence in models.

- The examples outlined in this report are included for the purposes of identifying the wider lessons for other public bodies looking to develop or work with statistical models and algorithms. It should be noted that these examples are therefore not exhaustive of all that was done in each country.
- In considering these case studies, we have drawn on the principles of the Code of Practice for Statistics . While not written explicitly to govern the use of statistical algorithms, the Code principles have underpinned how we gathered and evaluated evidence, namely:
  - **Trustworthiness:** the organisational context in which the model development took place, especially looking at transparency and openness
  - **Quality:** appropriate data and methods, and comprehensive quality assurance
  - **Value:** the extent to which the models served the public good.
- We considered the end-to-end processes, from receiving the direction from Ministers to awarding of grades and the planned appeals processes, rather than just the technical development of the algorithms themselves.
- We have undertaken this review using our regulatory framework, the Code of Practice for Statistics. The examples we have explored and findings we have reached should therefore not be taken to infer compliance or otherwise with any other regulatory or legal framework.

Throughout this report we have drawn on evidence from a number of sources including:

- Meetings with the qualification regulators
- Desk research of publicly available documents on the grade awarding process and the response to it
- Discussions with organisations who provide support in or have published guidance in the statistical models/ algorithm/ machine learning/ AI space
- Meetings with those conducting other reviews into the grade awarding process in 2020.

Report Structure

Part 1, Exploring the role of statistical models, outlines the role of models in supporting decision making. We define the terms that we use in the report, discuss model purposes and limitations and the guidance that is available.

Part 2, The grade awarding context, explores the grade awarding approaches taken in each country and the context in which they were developed. High level conclusions about the grade awarding process are presented.

Part 3, Learning from the grade awarding case studies, explores the approach to awarding grades in each of the four countries to identify wider lessons for other public bodies looking to develop or work with statistical models and algorithms.

Part 4, Commanding public confidence in statistical models, outlines our high-level findings on the grade awarding process and the wider lessons for government and other public bodies looking to develop statistical models to support decision, namely: public bodies developing statistical models, policy makers commissioning models and the centre of Government.

Recommendations for improvements to the support and guidance available to public bodies developing statistical models in the future are presented.

We would like to thank the qualification regulators, the other organisations that have helped us with this review and the members of our review Expert Oversight Group (list in Annex C).

Part 1 – Exploring the role of statistical models

This section of the report outlines the role of models in supporting decision making. We define the terms that we use in the report, discuss model purposes and limitations and guidance that is available.

1.1 Statistical models, algorithms and Artificial Intelligence

Throughout this report we have used the terms statistical model and algorithm when describing the various aspects of the models to deliver exam grades.

It should be noted, however, that terms such as statistical model, statistical algorithm, data-driven algorithms, machine learning, predictive analytics, automated decision making and artificial intelligence (AI), are frequently used interchangeably, often with different terms being used to describe the same process. In our view, the term AI, which currently has no universal accepted definition, is more and more being used as a catch all term to describe any form of data automation.

It should also be noted that the issue of terminology is not just specific to statistics, or even the world of government: these terms are currently used interchangeably all over the world in all sectors.

Given this, it is important to stress that the findings and recommendations in this report apply to any data-driven algorithm - whether derived from a statistical model, AI or machine-learning techniques or from human judgement.

1.2 Model functions and human involvement

There are two main types of models in existence: models where the output is the decision, and models where the output informs human decision-making. In models where the output is the decision there is simply no final human judgement involved and whatever decision the model has made becomes the final decision. In models where the output informs human decision-making the model itself is there to spot things which humans might miss but is not trusted in and of itself to make the final decision: that is left to human experts. This human involvement is known as having a ‘human in the loop’.

In both types of model there needs to be human involvement in the development of the model in selecting the data, setting the assumptions, coding the algorithm and in quality assurance. This can also be referred to as ‘human in the loop’. To avoid confusion we have used the terms ‘human verification’ and ‘human review’ in this report.

The type of model used depends almost solely on the context. In most public policy contexts, it is only appropriate for models to inform human decision-making, such as in medical diagnostics, but in other contexts, such as speech recognition or automated piloting, models function by providing the decision themselves.

Even though these are the two main overarching model types, there are, of course, other ways in which models can differ from each other. They can have:

- Different functions; some make predictions, some make identifications, some provide guidance and others simply make categorical decisions.
- Different forms; such as simple regression models, complex statistical models or machine learning models that add extra complexity and can be difficult to explain.
- Different derivations; such as use of historical data to model relationships or codification of previously human judgements
- Different deployments; as one-off models or be deployed to provide many outputs over time.

However they function, all the above would be referred to as statistical models. The conclusions of this report apply to all such models.

1.3 Model limitations

Modelled relationships between variables can be used to estimate the unknown value of one (or more) of those variables (for example, an exam grade) from known values for other variables (e.g. past performance). When cast in the form of a sequence of calculation instructions, this process is called an algorithm for estimating the unknown value. The statistical models used in the exam context are ‘data-driven’ models. That is, they are based purely on observed relationships between variables in data sets and on the statistical definitions of how those variables are constructed, and not on substantive psychological theory about how variables might be related.

Decisions are needed about what relationships should be included in the model. A simple model for estimating the time to drive between two points might be based on an assumption of constant speed, but a more sophisticated one might take likely traffic congestion into account. Observations from past data sets would reveal the extent and quality of data which could be obtained, and comparative evaluation of the models would inform the choice between them.

The accuracy of a statistical model, and corresponding algorithm, is dependent on a number of factors each of which identify potential limitations. These factors include how precisely the problem can be defined, what previous data are available, the amount of variability in the variable of interest, what assumptions have been made, and the quality of the data available to model the relationship. More detail on the limitations of algorithms and examples from the grade awarding scenario are in Annex A.

Returning to the example of a road journey between two points, defining the question might include questions such as how many breaks you want to take and whether you want to avoid any particular roads. Previous data might include extensive information on the roads you wish to use. Sources of variability might include the day or time of day when you are travelling and which roads you take. Assumptions might include the average speed that you can travel on particular roads and the time spent on breaks. The quality of the data may depend on the source; for example, people's recollections of doing the journey versus up-to-date mobile phone data. Separate models might then be developed for each road and for estimating the total break time. The overall algorithm would then be aggregated from these separate calculations.

1.4 Variability in the results of models

The results of statistical models are always subject to uncertainty. This uncertainty can be captured by giving a range or interval for predictions, rather than a single value. These ranges or intervals usually run either side of the central estimate. There are three main types of range that could be put around the output of a statistics-based algorithm

1. A **confidence interval**, which expresses the uncertainty in the estimate due to the limited sample size of the database. This can be made smaller by using a larger data set for the modelling. Confidence intervals are most important when data is sparse.
2. A **prediction interval**, which expresses the uncertainty of what might happen in the future in a specific instance. The prediction interval is a combination of a confidence interval for the uncertainty of a parameter, and the amount of variation of new observations about the value of that parameter. For example, our confidence interval might tell us that the average response is likely to lie between 70 and 80, while future individual values will have additional variation about that average. Prediction intervals are relevant only when making predictions.
3. A **discretionary interval**, which expresses reasonable leeway around a central estimate, to allow for individual and perhaps unquantifiable circumstances not taken into account in the algorithm. This may be based on subjective judgement. An example is the sentencing guidelines. This could be relevant when it is clear that the algorithm is limited and can provide only a ballpark figure.

1.5 Guidance and sources of information

A number of government departments and organisations are involved in developing and publishing guidance that touches on some of the learning points raised in this report.

We are concerned however, that the current landscape of guidance has created overlap, and therefore potential confusion, for public bodies looking to work with statistics models, algorithms and AI. In our view, there is a lack of coherence across the currently available guidance and therefore it is not as joined up as it should be.

A significant amount of guidance and best practice has been published in the last two years, particularly regarding AI, and new information continues to be released at regular intervals.

The Committee on Standards in Public Life stated, in its February 2020 report on Artificial Intelligence and Public Standards[2], the following:

> "Attempts to establish this governance and regulatory framework are emerging and developments are fast-moving. In the area of ethical principles and guidance, the Department for Culture, Media and Sport (DCMS), the Centre for Data Ethics and Innovation (CDEI) and the Office for AI have all published ethical principles for data-driven technology and AI. The Office for AI, the Government Digital Service (GDS), and the Alan Turing Institute have jointly issued A Guide to Using Artificial Intelligence in the Public Sector and draft guidelines on AI procurement. The Information Commissioner's Office (ICO) has also published its Auditing Framework for AI."

It is also worth noting that, due to the increased prevalence of AI as a catch-all term, there is a risk that valuable guidance and best practise in statistical modelling may be being overlooked due to how the guidance is being titled. This is particularly important for public bodies who are new to developing statistical models or algorithms.

Whilst it is not our intention for this report to be a comprehensive reference guide, we have provided in **Annex B** examples of recent publications related to the topic areas raised in the learning points of this report to illustrate the range of guidance available.

1.6 Summary

This section has explored the role of statistical models in supporting decision making. We have highlighted that:

- There is **confusion over terminology,** with various terms being used to describe data driven models and algorithms
- Models and algorithms can have a **variety of functions, forms and uses** but the conclusions from this review apply to all of them
- Data driven models have a **number of limitations** due to factors such as the quality of the input data and the assumptions made in developing the model
- The results of models are subject to **variability** and this should be recognised in how they are used
- There is a **fast-emerging community** developing and publishing guidance to support those developing statistical models, but this **guidance lacks coherence and leadership.**

In order to ensure public confidence in statistical models in the public sector, there needs to be improvements in the infrastructure to support them and a better understanding of how they can be used by those developing and commissioning them.

[2] Artificial Intelligence and Public Standards, A Review by the Committee on Standards in Public Life, February 2020 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF

Part 2 – The grade awarding context

This section of the report outlines the grade awarding approaches taken in each country and the context in which the statistical models were developed. Some key themes across the four countries are identified.

2.1 The approaches

The approaches to awarding grades were regulated by four bodies:

- In England, Office of Qualifications and Examinations Regulation (Ofqual)
- In Scotland, Scottish Qualifications Authority (SQA)
- In Wales, Qualifications Wales
- In Northern Ireland, Council for the Curriculum, Examinations & Assessment (CCEA).

The exact involvement of the qualification regulator in developing the approaches varied between jurisdictions, but broadly they were each directed by the respective governments to develop an approach to awarding grades in the absence of exams. They were instructed that approaches should award grades to students based on teacher assessments and other evidence, and that these should be standardised to maintain standards over time.

The approach taken to standardisation in each country differed. This review has not sought to draw conclusions on the suitability of the models themselves. However, to understand the implications of the approach to developing the models on public confidence it is helpful to understand the models at a high level. For the purposes of describing these models, we focus on the approach taken to awarding Highers and A-levels.

There were similarities and differences in the models used in each country to calculate grades. These arose from differences in the design of qualifications, the data that were available and the decisions taken by the qualification regulators and awarding organisations. In each country the qualification regulator or awarding organisation published an overview of their model, and the reason for it, in their technical documents on results day[3][4][5][6].

In all four countries the awarding organisations asked centres (schools or colleges) to provide an estimated grade for each student in each subject. Within grade they asked the centres to rank the candidates from the most likely to achieve the grade to the least likely. In addition, in Scotland centres were asked to place each student in one of 19 refined bands with 2 or 3 bands covering each grade.

In **England** a direct centre-level performance approach was used, whereby the previous distribution of grades achieved in the centre was adjusted for changes in the prior attainment of candidates in the current cohort compared with previous cohorts. The distribution of grades was then used to create the set of grades for that centre for 2020. These were allocated to students using the rank order supplied by the centre. Where there were a small number of students in a class or limited data, it was felt that the use of data would not be robust, and the estimated grades provided by the centres were used. A national level standardisation was then applied to ensure maintenance of standards overtime.

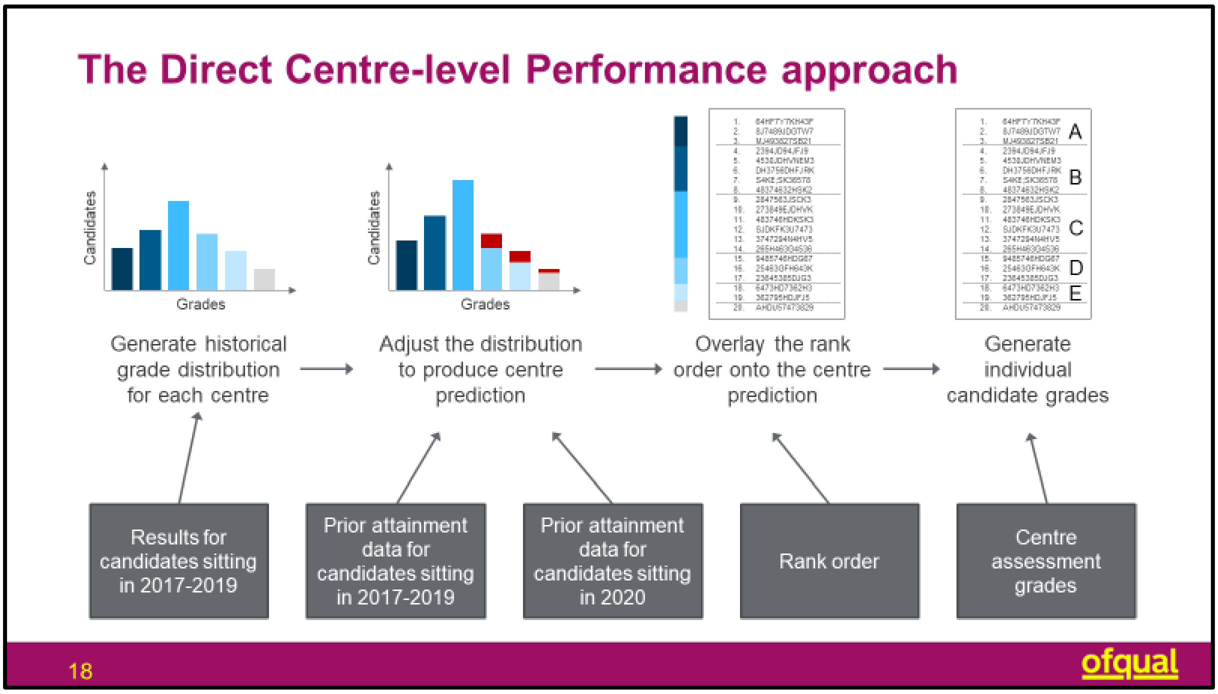
Figure 1 is a slide from the Ofqual Summer Symposium 2020 published on 21st July 2020 showing the Direct Centre-Level Performance Approach.

[3] CCEA –Compendium of approaches taken and formulae used in GCE and GCSE calculation of grades, Summer 2020
https://ccea.org.uk/downloads/docs/ccea-asset/General/Compendium%20of%20approaches%20taken%20and%20formulae%20used%20in%20GCE%20and%20GCSE%20calculation%20of%20grades%2C%20Summer%202020.pdf

[4] Ofqual - Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909368/6656-1_Awarding_GCSE__AS__A_level__advanced_extension_awards_and_extended_project_qualifications_in_summer_2020_-_interim_report.pdf

[5] SQA – Technical Report National Qualifications 2020 Awarding — Methodology Report
https://www.sqa.org.uk/sqa/files_ccc/SQAAwardingMethodology2020Report.pdf

[6] WJEC – Awarding grades for the June 2020 examination series: Qualifications Wales-regulated A-levels Methods report
https://www.wjec.co.uk/media/p0uiqio1/methods-report-qw-regulated-a-levels-1.pdf



The Direct Centre-level Performance approach

Generate historical grade distribution for each centre → Adjust the distribution to produce centre prediction → Overlay the rank order onto the centre prediction → Generate individual candidate grades

Results for candidates sitting in 2017-2019

Prior attainment data for candidates sitting in 2017-2019

Prior attainment data for candidates sitting in 2020

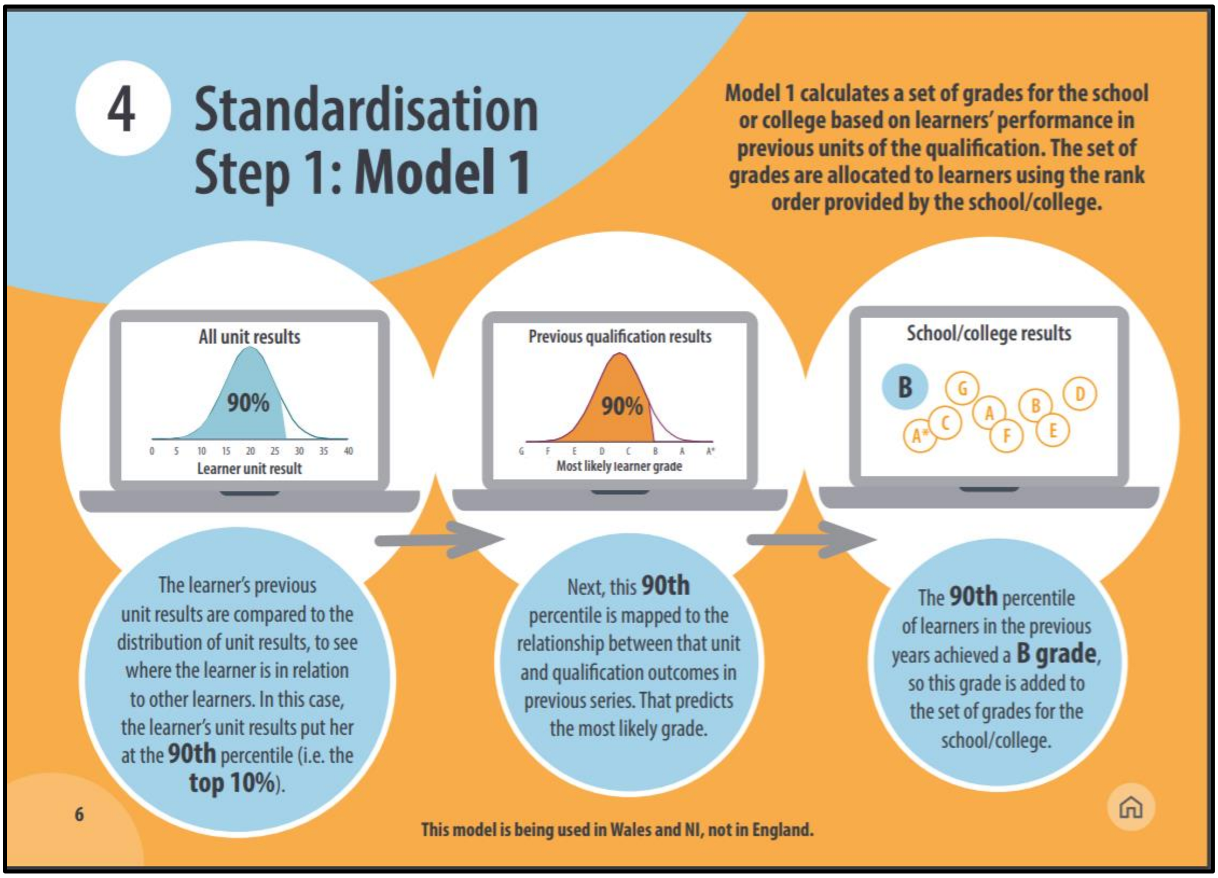
Rank order

Centre assessment grades

18

Figure 1 – Source: Ofqual Summer Symposium 2020 (PowerPoint) 21st July 2020

A-level qualifications in **Wales** are based on a unitised structure, with unit assessments available during the examination series each year. The student's previous unit results were compared to the distribution of unit results, to see where the student was in relation to other candidates and place them on the corresponding percentile. That percentile was mapped to the relationship between that unit and qualification outcomes in previous series. This approach predicted the most likely grade. This grade was added to the set of grades for the centre and these grades were then distributed to candidates according to the rank order of candidates provided by centres. A national-level standardisation was then applied to ensure maintenance of standards overtime. Figure 2 is a slide from the Qualifications Wales Summer 2020 results information pack showing the model used in step 1 of the approach to awarding unitised A-levels.



4 Standardisation Step 1: Model 1

Model 1 calculates a set of grades for the school or college based on learners' performance in previous units of the qualification. The set of grades are allocated to learners using the rank order provided by the school/college.

The learner's previous unit results are compared to the distribution of unit results, to see where the learner is in relation to other learners. In this case, the learner's unit results put her at the **90th** percentile (i.e. the **top 10%**).

Next, this **90th** percentile is mapped to the relationship between that unit and qualification outcomes in previous series. That predicts the most likely grade.

The **90th** percentile of learners in the previous years achieved a **B grade**, so this grade is added to the set of grades for the school/college.

This model is being used in Wales and NI, not in England.

Figure 2 – Source Qualifications Wales Summer 2020 results information pack

A-level qualifications in **Northern Ireland** are also based on a unitised structure. The approach drew on the methodology of estimating missing uniform marks in units for which a candidate has been absent. The process used a mean and standard deviation for the total marks in units the candidate had sat, which included an enhancement for re-sits, along with the mean and standard deviation for all units combined from the previous year to estimate a candidate's total A level uniform mark. A grade was then assigned using standard uniform mark grade boundaries. A centre's grade distribution was generated using these calculated grades and applied to the centre rank order.

The model in **Scotland** was based on moderation at centre level overlaid with national level constraints per subject to maintain standards. For each grade on each course at each centre, an allowable attainment range for 2020 was defined, against which a centre's estimated attainment was assessed. This allowable range was not rigidly constrained to historic attainment at the centre and included additional tolerances to allow for year-on-year variability in attainment. Where a centre's estimated attainment for a grade on a course was outside of the allowable range for that centre, it was brought within the allowable range by moving estimated refined bands using mathematical optimisation techniques. For example, if the estimated attainment for grade A was higher than the allowable range, those entries estimated by the centre to be in the lowest refined band(s) for grade A would then be moved to grade B. For each course, a national level check was undertaken to ensure that the cumulative attainment across all centres was within the pre-defined national constraints (which also had tolerances to allow for variability on historical national-level attainment).

2.2 Challenges of the grade awarding context

In reviewing the approach to developing these models to award grades it is important to recognise the context in which they were being developed. These included but are not limited to the following.

The qualification regulators and awarding organisations were required to develop and implement models in substantially shorter timescales and under more challenging circumstances than would normally happen for such projects.

The models needed to award the grades that students would most likely have received, whilst maintaining standards over time and not disadvantaging any groups. This created significant challenges for the qualification regulators.

Unlike in most situations when a model might be used to support decision making, the results had to be released on a single day that was determined in advance of the examinations being cancelled. In principle the statisticians could have stated that it was not possible to meet this date. In practice however, a fundamental aim was to ensure that the current cohort of students could continue to progress through education in a timely way without being adversely impacted. Short of an alternative solution to progression, outside of the remit of the qualification regulators, the grades were needed to enable that progression to happen.

Based on the results announced on the single of day of release, decisions were made that were difficult to change. In particular, on the days the A-level and Highers results were published decisions were made on university places and clearing was opened up for all other places. As a result, even if grades had been changed after this day, the student may not have been able to be awarded a place on their preferred course. Whilst this was outside of the remit of the qualification regulators, it is an important consequence of the grades they regulate. When schools were closed in March, university offers had been made on the basis of how results look in previous years. These offers were essentially contractually binding. When the results were re-issued based on centre assessed grades this caused issues for many universities.

There is variation in grades every year due to factors other than student ability. These factors include inconsistency in student performance between tests and occasions and variability in marking both between markers and for the same marker. Robust standardising and monitoring procedures are in place to reduce variability in marking across markers and within a marker’s allocation. The statistical techniques used to award grades in 2020 sought to overcome much of the variability in grades between centres and years, but there is always unpredictability that cannot be adjusted for, arising from these other factors. While the March 31 letter of direction from the Secretary of State for Education in England to Ofqual did not preclude the provision of a range of grades as the final output for each student, it is doubtful that this would have been acceptable. This meant that a single grade estimate for each student had to be awarded.

In light of the circumstances caused by the global pandemic which arose in 2020, the data available to develop the model was limited. In many scenarios where a statistical model is being developed to support decisions, the model is being used to replace the human element of weighing up a range of evidence. Therefore, the data available to that human is also available to build the model. In the grade awarding context, the main data usually available to determine the grade for the individual, marks from their assessments for that qualification, was not available. This created a challenge in identifying data that could be used in the model and be used to test the model.

The teacher estimated grades varied significantly from historic attainment for some centres in all countries for a variety of reasons. There was not sufficient time to verify the reasons for these differences.

2.3 Summary

This section has explored the approach to awarding grades in each of the four countries in the context of the challenges to developing the approach. We have identified a number of key themes across the four countries:

There can be an over-confidence in what statistical models can achieve: Statistical models have limitations. They are based on a number of assumptions and the data that is available. The results from statistical models are subject to variability. In the grade awarding context, the models were expected to predict a single grade for each individual on each course within constraints around maintaining standards and not disadvantaging any groups.

The task of awarding exam grades in the summer of 2020 was extremely difficult: It is important not to underestimate the scale of the task facing the qualification regulators and awarding organisations. They are small bodies and faced a shortage of both resources and time. Moreover, they did not have perfect information with which to work – for example, the practice of ranking students within a cohort was not one for which there was any historical evidence or experience. Finally, unlike most other uses of statistical models, they had to release their output on a single day to an entire cohort.

There were important differences in approach between the four parts of the UK: The biggest difference in approach was between Scotland the rest of the UK countries, as the model in Scotland made greater use of teacher estimated grades. Whilst similar in approach, there were also differences in England, Wales and Northern Ireland, for example around the use of teacher rankings, the extent of quality review and the use of prior performance of this cohort of students. The extent of differences between the four countries means that there is no single conclusion that can be drawn on the best approach.

It would be a mistake to think that the grade award decisions were made solely by a technical algorithm: There were many other aspects to the awards process, including public engagement, teacher judgement and appeals that were not the product of an algorithm. Grades were also determined by teacher judgements, whether in the form of rankings and/or centre assessed grades. In addition, the use of statistical models to support the setting and maintenance of standards at the cohort level is a common feature of awarding grades in a normal year, but what was required of the standardisation approach was very different in 2020.

It is for these reasons that this review has considered the end-to-end processes, from receiving the direction from Ministers to awarding of grades and the planned appeals processes, rather than just the technical development of the algorithms themselves.

Part 3 – Learning from the grade awarding experience

3.1 Introduction

In parts 1 and 2 of this report, we set out the role that statistical models have in decision-making and the specifics of the grade awarding process. In this section we explore the approach to awarding grades in each of the four countries in order to identify wider lessons for other public bodies looking to develop or work with statistical models and algorithms.

In identifying these lessons, we recognise the challenges and the trade-offs that had to be made in the grade awarding context. The pandemic meant the qualification regulators were working in a new and very challenging environment. What the grade awarding process has highlighted, however, is that achieving public confidence in a statistical model is complex and difficult.

Many of the decisions made supported public confidence. This review does not seek to make judgments on whether any individual qualification regulator performed well or badly. The examples in this section are included for the purposes of identifying the key learning and lessons for other public bodies looking to develop or work with statistical models and algorithms. We have drawn on evidence from several sources, including meeting with the qualification regulators and desk research of publicly available documents. These examples are not exhaustive of all that was done in each country. Where we have provided an example from one jurisdiction (England, Scotland, Wales or Northern Ireland) this should not be taken to mean that the example applies, or does not apply, in another jurisdiction. We have sought to balance the use of examples across the report.

We have undertaken this review using our regulatory framework, the Code of Practice for Statistics. As such, the examples we have explored, and findings we have reached, should therefore not be taken to infer compliance or otherwise with any other regulatory or legal framework.

We have identified that public confidence is supported by the following three key principles which we explore in more detail in this section.

Be open and trustworthy – ensuring transparency about the aims of the model and the model itself (including limitations), being open to and acting on feedback and ensuring the use of the model is ethical and legal.

Be rigorous and ensure quality throughout – establishing clear governance and accountability, involving the full range of subject matter and technical experts when developing the model and ensuring the data and outputs of the model are fully quality assured.

Meet the need and provide public value – engaging with commissioners of the model throughout, fully considering whether a model is the right approach, testing acceptability of the model with all affected groups and being clear on the timing and grounds for appeal against decisions supported by the model.

3.2 Key Principle – Be open and trustworthy

Statistical models can command public confidence when public bodies are open in how they develop and deploy them and are trustworthy. This section sets out the components that we have found to be key to being open and trustworthy in the approach to developing and deploying models.

3.2.1 Public bodies must be open in the development of models

3.2.1.1 Transparency of aims

Transparency of aims is about ensuring that it is clear what you are trying to achieve and whether it is achievable. In the case of models that support decision making in the public sector, it is also about stating these aims publicly so it is clear to those affected by the model what it is trying to achieve.

Examples from the 2020 grade awarding processes

The qualification regulators were given a direction from the relevant minister or Secretary of State. This set out the Government's view on what the objective of the qualification regulators' approach should be and in the case of Northern Ireland the Department of Education (Northern Ireland)'s instruction on how this should be achieved. The qualification regulators sought to further clarify the objective by publishing the aims or principles of the standardisation process. Both Ofqual and Qualifications Wales consulted externally. Although the task in each country was similar, the aims and principles differed in their wording and in what was included. Table 1 lists the aims or principles published by each qualification regulator.

Table 1: Aims or principles published by each qualification regulator

Ofqual (England) • to provide students with the grades that they would most likely have achieved had they been able to complete their assessments in summer 2020 • to apply a common standardisation approach, within and across subjects, for as many students as possible • to use a method that is transparent and easy to explain, wherever possible, to encourage engagement and build confidence • to protect, so far as is possible, all students from being systematically advantaged or disadvantaged, notwithstanding their socio-economic background or whether they have a protected characteristic • to be deliverable by exam boards in a consistent and timely way that they can quality assure and can be overseen effectively by Ofqual	**Qualifications Wales (Wales)** • Aim 1: Learners for whom a qualification level centre assessment grade and rank order are submitted will receive a grade. • Aim2: National outcomes will be broadly similar to those in previous years to reduce the risk of unfairness for learners over time and maintain public confidence. • Aim 3: As far as possible, the process for awarding grades will not systematically advantage or disadvantage learners, including those with characteristics protected by equalities legislation. • Aim 4: The statistical standardisation model will use a range of evidence to calculate the likely grades that learners would have achieved, had they been able to complete their assessments
Scottish Qualifications Authority (Scotland) • Fairness to all learners • Safe and secure certification of our qualifications, while following the latest public health advice • Maintaining the integrity and credibility of our qualifications system, ensuring that standards are maintained over time, in the interests of learners	**Council for the Curriculum, Examinations & Assessment (Northern Ireland)** • To award grades to candidates that they would have most likely achieved had they been able to complete their assessments in summer 2020. • To award grades to all candidates using as far as possible a consistent approach for all candidates in all qualifications. • To ensure the method used to calculate grades in summer 2020 is fair to all candidates. • To ensure that outcomes are as similar as possible to previous years and so maintain standards. • To provide an explanation that can be understood by the public on the approach used in summer 2020 to award grades.

The qualification regulators recognised that it was not an easy task. For example, Qualifications Wales highlighted in its 1 May blog that it was “a totally unique situation within an unprecedented time in living memory for all.” and, “The decisions we have to take in these difficult circumstances won’t always be popular with everyone.”

The difficulty of using a statistical model to award grades was also recognised by Ofqual. As set out in the written statement from their Chair to the Education Select Committee, their initial advice to the Secretary of State was to recommend other approaches to awarding qualifications rather than calculating grades.

The aims placed constraints on the model that could be chosen and the results from the model. For example, Qualifications Wales reported to us that their aim - that national results had to be broadly similar to previous years - constrained the degree to which the approach could lead to grade inflation. At the same time, the approach needed to ensure that no groups were systemically disadvantaged.

CCEA and Qualifications Wales reported to us the measures they put in place to ensure that their aims were met.

In each country there was more than one aim or principle that needed to be met. This meant that tradeoffs may have had to be made where there were tensions between the aims. This was recognised, for example by Ofqual in their consultation document, where they state, “Where the aims listed above are in tension (for example, accuracy of approach versus ease of explanation), we will seek to find an optimal balance.” In addition, The Wales Independent Review interim report found that in developing the approach to calculating grades, the principle of maintaining confidence in and credibility of the qualifications system appeared to be prioritised over the other aims.

Lessons for others developing models

The qualification regulators attempted to set out clearly the aims and principles of the approach. Despite this, the overall aims and principles may have been in tension.

In our view, it was not clear whether all of the published aims or principles could be achieved at the same time with a statistical model or that all of the regulators referred back to the aims to ensure that they had been met fully.

We believe the potential tensions between the aims, and likely consequences of these, may not have been clear to those impacted by the model.

Others developing models should:

- Ensure aims are clearly stated, achievable with a statistical model and accepted publicly**.**
- When developing and evaluating a model, regularly assess that the stated aims will be delivered.
- Be clear where there are tensions between aims and what the relative priorities are.

3.2.1.2 Openness to feedback during model development

Openness to feedback is about engaging with a broad range of stakeholders throughout the model development and listening and responding to their feedback. This helps public bodies to address potential issues and gain support for the broad approach.

Examples from the 2020 grade awarding processes

All of the qualification regulators sought to gain feedback on their aims and principles.

For example**,** Ofqual and Qualifications Wales held public consultations. Reports on the feedback to the consultation and decisions made on the basis of the feedback were published by Qualifications Wales and Ofqual. SQA and CCEA did not hold formal public consultations but did seek feedback from established groups and groups representing students. CCEA reported to us that general qualifications policy is owned by Department of Education (Northern Ireland), so CCEA would have had to have instruction directly from the minister in order to consult publicly on the principles of standardisation.

Whilst all the qualification regulators sought feedback on their aims, it was reported to us that it was not always clear what actions they could take based on it. For example, concerns were raised during the Qualifications Wales consultation around the fairness for individuals this year given the unprecedented circumstances and the use of centre-level data. Qualifications Wales told us that although concerns were raised, alternative statistical solutions to standardisation were not offered. This made it difficult for them to make changes based on the feedback.

Lessons for others developing models

The qualification regulators sought feedback on their aims and principles.

It is not clear to us whether it was always possible for concerns that were raised to be adequately resolved given the time and resource constraints for the model development. However, we conclude that their approach was reasonable in the circumstances.

Others developing models should:

- Ensure that continuous engagement is sought, and feedback is acted on throughout the model development.

3.2.2 Public bodies must be open in the deployment of models

3.2.2.1 Transparency of the model and limitations

Transparency of the model and its limitations involves ensuring it is communicated in a way that is understandable and meaningful to the intended audience. Public confidence requires both technical experts to be able to evaluate it, and lay people to be able to engage with and understand the essentials and the likely outcome for them.

Examples from the 2020 grade awarding processes

On results day methodology guidance on how the grades had been calculated and alternative methods considered were published in all the countries. These documents provided a level of detail that enabled technical experts to gain an overall understanding of the approaches that were considered. They were not accessible to non-statisticians, but they were not designed to be. Ofqual have since outlined in a December 2020 blog that they are working towards making the underlying data available to researchers to enable them to replicate and evaluate the methods in line with their statement at the Education Select Committee. Ofqual made the code that they developed for the awarding organisations available on GitHub in December 2020.

All the qualification regulators and awarding organisations undertook extensive communication activities. These ranged from social media advertising campaigns to targeted communication for those affected by the model. Included in this were YouTube videos aimed at students, teachers and parents published by Ofqual, Qualifications Wales, SQA and CCEA. These helped explain the high-level method in a simple way. In our view, however, there was potential for different people to have different interpretations of how the centre assessed grades would be used in the model from these videos. For example, the Ofqual YouTube video stated that the model would compare centre assessed grades with the centre’s historic grades and adjust grades where the centre assessed grades appear more severe or generous. Similarly, in the CCEA YouTube videos, there was significant discussion of the centre assessed grades. This may have given the impression that centre assessed grades would have a larger role in the model than they ultimately did.

In addition, SQA had an online FAQ for students, parents and teachers and CCEA provided interviews to journalists and local radio broadcasts to help communicate how grades would be calculated. Ofqual held a Summer Symposium and Qualifications Wales held a webinar to explain their approach to stakeholders.

There was limited information in these communications around how the data provided by schools and other centres would be used in the model. We appreciate that this was in part due to the challenging timelines for development of the model, resulting in decisions around the model not being decided until after the centre assessment data had been requested and received. In addition, Ofqual told us that had they provided information earlier, it would have negatively impacted the quality of data from schools.

Lessons for others developing models

The qualification regulators undertook a number of activities to communicate information about the model to technical experts and those affected by the model.

The need to communicate about the model whilst also developing it inevitably made it difficult to be fully transparent about the models prior to results day. In particular, there was limited information about the models and their limitations available for technical experts, those supplying the input data and those affected by the model to help them understand it.

In our view, as a result there was limited public understanding or awareness of these limitations prior to results day.

Others developing models should:

- Clearly explain the limitations of algorithms and the approach adopted to all stakeholders.
- Make data available to researchers to help evaluate methods.
- Be clear if the final model is not known when requesting the input data or engaging with stakeholder groups.
- Be clear what aspects of the model have been tested.
- Be transparent about the limitations of the available data.

3.2.2.2 Transparency of the process being replaced by a model

Transparency of the process being replaced by a model is about ensuring a level of familiarity with the existing process.

Most statistical models will be introduced into public services where there is already an existing process. Where this existing process is not well understood, it is possible that the statistical models will highlight existing issues, or even be blamed for things which are inherent in the process.

Examples from the 2020 grade awarding processes

There is variability in examination grades in a normal year and qualifications regulators work with awarding organisations to standardise these in the interests of fairness. It is widely assumed that a student sits an exam and the examiner gives the student a mark and grade based on that paper resulting in the final grade. As detailed in Annex A, there are many sources of variability and processes in place to reduce them. In essence, the student’s mark can differ from what was predicted if they perform poorly or very well on the day of the exam.

In addition, the use of statistical evidence and expert judgments to support the setting and maintenance of standards at the cohort level and ensure inter-board comparability is a common feature of awarding grades in a normal year. These approaches may not be well understood in general. Nevertheless, as well-established processes, based on expert judgments, they are able to command public confidence.

Whilst this information is generally publicly available, none of the regulators explicitly drew attention to the details of the methods normally used to estimate or standardise grades or reduce variability in marking when communicating the models.

Lessons for others developing models

In our opinion, notwithstanding the very extensive work to raise awareness, there is general limited understanding amongst students and parents about the sources of variability in examination grades in a normal year and the processes used to reduce them. As a result, when the unfamiliar 2020 approach was presented publicly, people may have assumed that an entirely new, machine-led approach was being introduced, and this may have raised their concerns.

An understanding of these processes might have helped parents and students to better understand the variability in the normal process and the potential impact of standardisation on the 2020 results. This lack of understanding may have made it more difficult for the models to command public confidence.

This issue of broader understanding would have been very hard for the regulators to address in the time available.

Others developing models should:

- When using an algorithm in place of another process, communicate the strengths and limitations of the usual process so that all audiences can understand the usual level of uncertainty.

3.2.2.3 Understanding of the impact of social inequalities in the input data

Statistical models and algorithms do not create inequalities in their own right. Differences in outcomes between groups arise from historical patterns in the data and the assumptions that the model is based on. The potential impact arising from patterns in the input data and assumptions in the model, and how it will be reduced, should be made clear. Otherwise, the model will not command public confidence as it will be seen to include, and therefore create inequalities in the results

Examples from the 2020 grade awarding processes

In all four countries the previous history of grades at the centre was a major input to the grades that the students of 2020 at that centre would receive for at least some of their qualifications. For example, in England the initial distribution of grades for a centre was based on analysis of that centre’s grades over the past three years and also included information about the prior attainment of candidates in that centre’s cohort, where available. Where the model used centre level prior performance and centres had historically low grades, this data may have led to these centres receiving low initial grades in the results of the model.

In Scotland the tolerances used to identify which centre’s grades needed to be standardised were calculated based on the prior performance of each centre. Although the tolerances allowed for some improvement, the range of grades within the tolerances would have been lower for a school with lower historic performance than for a school with higher historic performance. In all the centre level models there was an assumption that a centre that has previously had low grades would have low grades in 2020 and similarly a centre that had high grades would have high grades in 2020. This made it difficult for students at these centres to be awarded higher or lower grades than those previously achieved at that centre using the algorithm.

In Wales and Northern Ireland, the A-level model was not based on the performance of prior cohorts of students at the centre but used the AS level results of the 2020 cohort to model the distribution of grades for the centre. However, the planned models for AS and GCSE did include the performance of prior cohorts, with the exception of some unitised GCSEs in Wales which used the same approach as for A-level.

Historic results show that pupils with some characteristics have lower attainment and this may have led to the prior performance for centres with a higher proportion of students with these characteristics being lower and hence the distribution of grades for that centre from the model being lower.

Differences in results are not in themselves a bias. Grades aim to measure and represent different levels of attainment. If they reflect genuinely different levels of attainment in the population then this is not a bias within the data, but may reflect broader socioeconomic factors. Indeed, qualification results in past years have played an important role in highlighting and evidencing social inequality.

Nevertheless, before the results were released, concerns were raised publicly about the impact of the models on disadvantaged groups. For example, the Education Select Committee produced a report in July Getting the grades they've earned. Covid-19: the cancellation of exams and 'calculated' grades which detailed some of their concerns around the model potentially disadvantaging categories of students including Black, Asian and Minority Ethnic (BAME) pupils, pupils with special educational needs and disability (SEND), looked after children, and free school meal (FSM) eligible pupils.

The regulators recognised that their approach would mean that patterns of attainment in past years, including any inequalities, would be replicated. This concept was discussed extensively within the governance and oversight structures. All the regulators carried out and published a variety of equality impact analyses on potentially disadvantaged categories of students at an aggregate level prior to or on results day. These analyses were based on the premise that attainment gaps should not widen, and the analyses did indeed confirm that gaps had not widened.

Further detailed inequalities analyses were also published after the results were released. For example, in November 2020 Ofqual published detailed analysis of CAGs and calculated grades by student characteristics and, in December 2020, published analysis of calculated and final grades by centre type.

Lessons for others developing models

All the regulators carried out and published a variety of equality impact analyses on potentially disadvantaged categories of students at an aggregate level.

But there was limited public discussion ahead of the release of results about the likely historical patterns in the underlying data and how they would impact on the results from the model. When results were released, there was widespread public perception that students in lower socio-economic groups were more likely to have lower grades compared with the centre assessed grades than those in higher socio-economic groups which caused a significant degree of public concern. The regulators provided analysis to provide assurance on these concerns, but in some cases, only some time after the perception itself had arisen.

Others developing models should:

- Be clear about the social inequalities that exist in the underlying data and how they have been treated.

3.2.3 Public bodies are trustworthy

3.2.3.1 Honesty and integrity

Honesty and integrity are paramount when considering public confidence and are central to the trustworthiness of any organisation. For teams developing statistical models to support decisions, this means being honest about the approach they are taking and the reasons for it.

The public must have confidence in the people developing statistical models to support decisions in order for the models to command confidence.

Examples from the 2020 grade awarding processes

In our view the teams in all of the qualification regulators and awarding bodies acted with honesty and integrity. Throughout this review the teams were honest with us about the approach they took in developing and deploying the models. They willingly met with us and sent us information to inform the review, whilst also being under pressure from a number of other sources.

Honesty and integrity also operate at the organisational level. In the case of the qualification regulators their purpose includes promoting public confidence in the qualifications that they regulate.

Lessons for others developing models

In our view, the teams in all of the qualification regulators and awarding bodies acted with honesty and integrity. All were trying to develop models that would provide students with the most accurate grade and enable them to progress through the education system.

Conflicts may have existed between the core purpose of promoting confidence in the qualifications system and being transparent about the limitations of the approach. However, we found no evidence that this had an impact on transparency in practice.

Others developing models should:

- Be aware of and, if necessary, address any organisational objectives which might impact on the development and communication of a model.

3.2.3.2 Transparency and communication

It is important to recognise that transparency and communication are not the same thing. Communication is about imparting the information which is required to make someone understand. Transparency implies openness and accountability as well as communication. Without transparency, public bodies will not be trustworthy and statistical models developed by them are less likely to command public confidence.

Examples from the 2020 grade awarding processes

All of the regulators used a variety of mechanisms to ensure that they were communicating with teachers, students and parents around the models for calculating grades. Ofqual used communications experts to develop a communication strategy and segment their audience using different communication channels and social media. Their approach had several objectives in mind – to reassure students, parents, teachers and employers of its fairness and comparability with other years, to ensure that there were ‘no surprises’ and manage expectations and to project a message of unity. CCEA also developed a communication strategy at the start of the process.

There were concerns around the impact on input data if details of the model were available prior to centre assessment data being submitted. There were also concerns about the potential for people to try to calculate the grades if more details of the model were available before results day.

SQA told us they did not publicly announce prior to results day that some of the centre estimates were significantly above historic attainment, as they did not want to cause undue anxiety, with students worrying that their grades might be one of the ones affected. In Wales there was limited transparency and public engagement around the model and its potential limitations. In part this was an active decision as Qualifications Wales were concerned that highlighting the limitations of the model may undermine confidence in the qualifications system and the credibility of the standardised grades. In addition, Qualifications Wales highlighted that communication of the technical concepts required the limited resource of the people developing the model and further work here would have impacted on their ability to deliver the model.

Whilst these are reasonable concerns, limited transparency around the models contributed to the lack of public confidence. For example, the Rapid Review of National Qualifications Experience 2020 in Scotland found that “There is widespread criticism by respondents of SQA for a perceived lack of transparency and a failure to engage in participative development of solutions with stakeholders.”

Both the Education Select Committee in England and the Scottish Parliament requested that further information be made available. The Education Select Committee asked Ofqual in July 2020 to be completely transparent about its standardisation model and to publish it to allow time for scrutiny.

Lessons for others developing models

In our view, there was a very significant emphasis on communication by all the regulators.

Whilst the regulators put resource into developing communication strategies and considering how to communicate with the different audiences, they were not publicly perceived as being transparent. This gap between their very significant efforts and the public perception shows the scale of the communication challenge, and indicates that, for other organisations, the public communication challenge should not be under-estimated.

Others developing models should:

- Focus on transparency as well as communication throughout the development of a model.
- Ensure that communication strategies support all affected groups to understand the impact of a model.

3.2.3.3 Ethical use of data and models

A model will not command public confidence unless complete consideration of the ethical use of it has been made. Ethical use of data and models requires an organisation to reflect on both whether it is possible and if so, whether it is the right thing to do. There are legal frameworks to support organisations to ensure that the ethics of their approach is considered. For example, there are specific provisions in data protection legislation about the use of automated decision making. Whilst adherence to these provisions does not guarantee that the use of an algorithm will be considered ethical, it will ensure that it is legal.

Central to any discussion on the ethical use of data and models to support decisions about individuals will be whether the rights of those individuals have been complied with. In the exams scenario the UN Convention on the Rights of the Child was relevant. This sets out that public bodies should consider the best interests of the child when doing anything that affects children.

There is also guidance and toolkits available to support organisations in considering the ethics of their approach when using data driven models - for example, European Commission Expert Group: Ethics guidelines for trustworthy AI , ICO Guidance on AI and Data Protection and the National Statistics Data Ethics Advisory Committee toolkit.

Examples from the 2020 grade awarding processes

The qualification regulators were clearly aware of their legal responsibilities and reported to us that they took legal advice about possible approaches - for example, around whether to seek the input of schools where there were large changes from the centre assessed grades.

All the qualification regulators undertook a number of activities to explore the impact of their solutions on specific groups to ensure fairness. These activities included a variety of impact assessments and detailed equalities analyses.

Our review of the available guidance on ethics has highlighted that, although a lot of guidance has been produced recently, the current landscape of guidance has created overlap, and therefore potential confusion, for public bodies looking to work with statistical models, algorithms and AI. As such, the qualification regulators may not have been fully aware of the all of guidance around the ethical use of data and models beyond that relating to impact assessments.

Lessons for others developing models

In our view, the use of impact assessments by the qualification regulators helped them consider their legal duties and the impact of the models on specific groups.

The available guidance around the ethics of data and models/ AI/ algorithms may not have been recognised as being relevant. Better signposting to guidance to support public bodies, such as that in Annex B, may have helped, particularly in the exam context, where timescales were tight and the qualification regulators were seeking a solution outside of their business as usual.

Others developing models should:

- Identify relevant guidance around the ethical use of data and models to ensure that both legal and ethical issues are considered.

3.2.3.4 Involvement of lead analyst

The Head of Profession for Statistics, Chief Statistician in the devolved administrations or lead statistician has a key role in upholding and advocating for trustworthiness, quality and value to all those involved in producing, publishing and using statistics and data in the organisation. We believe that complying with these pillars of the Code of Practice for Statistics can also help ensure public confidence in statistical models.

Examples from the 2020 grade awarding processes

In Wales, Scotland and Northern Ireland, the lead statistician took a key role in the governance and oversight. This enabled key elements of the Code to be followed and advocated for, including trustworthiness. For example, the lead statistician at Qualifications Wales took ownership of the process for providing analytical support for their Board’s governance of the models which were developed and tested by WJEC. The lead statistician advocated for many principles in the Code to support the trustworthiness of the approach.

Lessons for others developing models

In our view, where the lead statistician was involved in the development and deployment of the models, the trustworthiness, quality and value of the approach was clearly considered.

We recognise that it will not always be statisticians that develop statistical models. Central government are moving to lead analysts having responsibility for advocating for the Code of Practice. Ensuring that the lead analyst has a role within the governance of statistical models can help to ensure that trustworthiness, quality and value are considered, which in turn should lead to improved public confidence. This can be supported by consulting the National Statistician where advance statistical techniques are being used.

Others developing models should:

- Ensure the lead analyst has a role within the governance to support trustworthiness, quality and value.
- Ensure that where advanced statistical techniques are being considered the National Statistician should be consulted the for advice and guidance.

3.3 Key Principle – Be rigorous and ensure quality throughout

Statistical models can command public confidence when public bodies are **rigorous** in how they develop them and apply them and **ensure quality throughout**. This section sets out the components that we have found to be key to being rigorous in the development and application of statistical models and ensuring quality throughout**.**

3.3.1 Public bodies must be rigorous in how they develop models

3.3.1.1 Involvement of analytical and subject matter experts

When developing models, public bodies should include both analytical and subject matter experts. The subject matter experts bring an understanding of the context in which the model will be deployed. Analytical experts will bring the theoretical expertise and an ability to test the approach against a wider range of contexts.

This is particularly important where there is no settled professional consensus or best practice. For the production of economic statistics, for example, there is clear international guidance. Similarly, there is clear best practice on how to carry out surveys. The extent of professional consensus is therefore an important indicator of the degree of risk involved in the development of models, and the challenge in securing public confidence.

Examples from the 2020 grade awarding processes

The grades context was a novel issue. There was no existing guidance or best practice on the best way to calculate grades in the absence of exams.

There was a great deal of collaboration within the exams system to develop the models. For example, there were regular meetings between the regulators and awarding organisations and between countries – particularly England, Wales and Northern Ireland. This collaboration enabled solutions to be found within the timeframes as different organisations took on responsibility for analysing different solutions.

Ofqual set up an External Advisory Group (EAG) to consult on decisions around the model. This group was made up of academics and statistical specialists in qualifications and education. Qualifications Wales did not seek separate professional advice beyond WJEC, the other qualification regulators and Welsh Government education statisticians, but drew on the advice that Ofqual had received.

SQA appointed two private consultants to help them to both develop and test their model, as well as to provide critique and challenge. SQA methodology was reviewed by subject experts within SQA.

CCEA conducted peer reviews of their model with a private contractor and universities.

While the qualification regulators were drawing on the extensive community of expertise in designing examination systems, concern was expressed publicly about a lack collaboration with experts outside of the education community to help develop the solution. As was widely reported in the media shortly after results day in England, the Royal Statistical Society offered specific fellows to help but this was not taken up, although the RSS and Ofqual have disagreed publicly about details of the circumstances.

Late in the process Ofqual sought advice from the Office for National Statistics' Methodology Advisory Service, who specialise in methods relating to production of official statistics. An initial review was completed based on the technical documentation, much of which became publicly available on results day, and advice was provided to Ofqual with the aim of being helpful should a similar approach need to be used again in the future. This included that they should seek a wider and independent review of the standardisation process if developing a similar model in future and that they should make the student and centre-level data from 2020 available to researchers. As highlighted earlier in this report, Ofqual have told us that they are working towards making sharing data via the Office for National Statistics secure research service.

Lessons for others developing models

The qualification regulators drew on extensive expertise within the qualifications and education context.

Despite this, there was limited professional statistical consensus on the proposed method. The methods were not exposed to the widest possible audience of analytical and subject matter experts, though we acknowledge that time constraints were a limiting factor in this case.

There was not an obvious group within or outside government that the qualification regulators could approach to help the support development of the model and gain statistical, professional consensus. The Methodology Advisory Service and Data Science Campus could have provided timely support had they been contacted earlier in the process of developing the models. In this environment of limited consensus, it would have helped had the regulators adopted more formal Red Team models of challenge, or organised independent advisory panels drawn from a wider range of professional perspectives.

In our view, this limited professional consensus, and the public discussion of it, is likely to have undermined public confidence in the models.

Others developing models should:

- Where there is no national or international guidance or established best practice, it is important to ascertain whether there is a professional consensus on the proposed method.
- Methods should be exposed to the widest possible professional audience of both analytical and subject matter experts, including the potential for external challenge through Red Teams or independent advisory panels.

3.3.1.2 Decisions on key concepts within the model

As stated by the Royal Statistical Society in their statement on grade adjustment in 2020 examinations in the UK: “Any statistical algorithm embeds a range of judgements and choices; it is not simply a technically obvious and neutral procedure.” The culmination of all those decisions gives the overall model. The face validity of that overall model impacts heavily on public confidence. If the model does not appear valid then there will not be confidence in it.

Examples from the 2020 grade awarding processes

Prior to developing and testing specific models, Ofqual took a series of high-level decisions around the approach. These included the data they would use, the level of standardisation and the starting point of the standardisation model. These were discussed and reviewed with their External Advisory Group. These choices provided them with clarity on the overall approach and enabled them to identify which models to test. Some of the alternative models suggested by others did not fit with that approach and so had already been discounted.

In England, Wales and Northern Ireland the decision was made to place less weight on the centre assessed grades in the model. These decisions were based on accepted research evidence that teachers are better at relative judgements on performance than absolute ones and analysis of historic predicted grades.

The centre ranking data played a large part in the individual grades received by students. However, rankings are also subject to uncertainty. In order to explore the impact of the uncertainty in ranking data, WJEC carried out analysis of the impact on accuracy of adding a random error to the true rank from 2019. To allow for the uncertainty in rankings, they permitted tied rankings within the centre assessed data. In England and Northern Ireland, tied rankings were not permitted, except for in large centres in England.

For centres with a small entry in a subject, the centre assessed grades (CAGs) were used as the final result. This was due to there being insufficient historical data to use the Direct Centre Performance approach. This was a source of public concern as the model was not perceived to be treating all students consistently. Qualifications Wales and Ofqual reported to us that they considered the use of alternative methods such as combining centres and randomised elements in the model, but these were deemed not to be acceptable to centres and learners.

As A-levels in Wales and Northern Ireland are unitised qualifications, WJEC and CCEA had the marks of students in completed AS qualifications. These were used to predict a final grade for each candidate. This grade was added to the set of grades for the centre and allocated using teacher rankings. They could have been used directly for those candidates instead of using the two-step algorithm. The rationale for not doing this was that the rankings provided more up to date information on the performance of the candidates.

In Scotland, the technical document states that they aim to “move the minimum number of entries.” However, refined bands (1-19) were used when adjusting the grades in Scotland, resulting in large groups of entries being moved. SQA told us that this was compounded by “bulges” at the bottom of the refined bands.

The Rapid review of national qualifications experience 2020 in Scotland found that a potential problem “might be the way in which the optimisation problem was defined.” This focussed around whether the program aimed to prevent large movements in grades or minimise them. SQA reported to us that they were not trying to prevent these movements but to minimise them.

It should be recognised that the data available to the qualification regulators on which to develop the models was limited. The centre assessment data, both grades and rankings, had not previously been provided in this format and set of circumstances. Therefore, models based on them could not be tested using prior data. This resulted in models that relied on past performance of centres. It also meant that the full model could not be tested.

In developing the models, the regulators and awarding organisations tested the possible approaches against a selection of subjects. Whilst we recognise that time and resource constraints would have meant that the approaches could not have been tested against all subjects, the rationale for the choice of subjects and the robustness of decisions on the particular set of subjects was not described fully publicly. Analyses were not performed at student level to test the impact on individuals. We recognise, however, that this may not have been possible in this context.

Lessons for others developing models

The qualification regulators and awarding organisations carried out extensive analysis in order to make decisions about the key concepts in the model. Most of the individual decisions seemed valid in their own right. Despite this, the validity of the overall model was not clear to those affected and this impacted on public confidence.

The teams developing the models faced a challenge in that there was limited data available on which to develop and test the models. These limitations drove some of the decisions on the model. The limitations of the data resulted in a different approach for small centres in England, Wales and Northern Ireland. In our opinion the impact of this on perceived fairness undermined confidence in the models.

In addition, the models could not be developed and tested on the data that was used to run the model, nor could they be tested at the individual level.

Others developing models should:

- Be clear how the limitations of the available data impact on decisions on key concepts in a model.
- Ensure that a consistent approach is applied to all individuals when making key decisions about a model.
- Develop and test the models on the data that will be used to run it.
- Assess a model against the groupings that will be affected by the output and those of interest to the public (where these are different).

3.3.2 Public bodies must be rigorous in how they apply models

3.3.2.1 Clear governance and accountability

As technology enables us to create ever more sophisticated models using ever larger and more complex datasets, the need for effective governance and oversight of them becomes even more important. Where the governance or accountability is not clear it is less likely that the model will command public confidence. The Aqua Book: guidance on producing quality analysis for government states that ‘It is important that departments and agencies have a cascade of accountability and responsibility from their senior management teams down throughout their organisation’.

Examples from the 2020 grade awarding processes

All the organisations had key governance and sign off processes internally and the risks around the model were raised at board level in each country.

Ofqual documented clearly where decisions would be taken at board level and where decisions would be delegated to the analysts. All of these board papers are now published and this shows how the governance operated as well as how decisions were taken when deciding on which standardisation model to adopt. Ofqual operated with a number of governance groups including the Standards and Technical Issues Group (STIG), the chief regulator, a policy and implementation group and an External Advisory Group on Exam Grading.

SQA had several bodies available to assist them. These included their Qualifications Committee, Advisory Group and Code of Practice Governance Group. SQA were also able to make use of Scottish Government’s Qualifications Contingency Group, which contains already established key stakeholders who SQA used to consult on key documents.

As is the case in all years, CCEA and Qualifications Wales collaborated with governance groups in England, such the STIG, as well as having clear governance arrangements in their own countries.

Ofqual and Qualifications Wales included information on limitations in papers to their boards. In Scotland, a research and evidence report set out different technical options, including their strengths and limitations. This report was shared with boards and later formed part of the published Technical Report.

Lessons for others developing models

In our view, all the qualification regulators put in place clear governance structures and attempted to make sure it was clear where decisions would be made.

It is not clear to us however how risks around public acceptability and what was achievable in the timescales were effectively managed within these governance structures.

Others developing models should:

- Ensure the governance and accountability are clear.
- Ensure that the limitations of statistical models are clearly communicated to non-statisticians within the governance structure.
- Ensure that risks around public acceptability and what is achievable with a model are managed within the governance structure.

3.3.3 Public bodies must ensure quality throughout

3.3.3.1 Agreed and documented quality criteria

As set out in the Code, statistics should be produced to a level of quality that meets users' needs. The Aqua book specifies that 'Quality assurance considerations should be taken into account throughout the life cycle of the analysis and not just at the end.' In order to command public confidence that level of quality needs to be clearly set out and agreed in advance.

Examples from the 2020 grade awarding processes

In their methods report, WJEC set out the key principles which they used to select the model. These were the key principles of assessment – validity, reliability, fairness, manageability, and comparability. Prior to WJEC providing recommended approaches, Qualifications Wales undertook a range of analyses themselves to familiarise themselves with the data, methods, and likely accuracy they could expect. They shared relevant analyses with WJEC so that these could inform their recommended approaches. In general, findings influenced their view that national results should be slightly higher than in recent years. Qualifications Wales undertook analysis to examine how well standard statistical techniques such as linear regression could predict results. This gave them an indication of the likely 'best' prediction accuracy to compare potential models against.

Ofqual highlighted to us that published research literature shows that only around 50% of teacher estimates match the final grade awarded and that around 10% of grades will differ from teacher estimates by two or more grades. There was therefore an expectation that at as many as half of the grades could differ from the centre assessed grades and there would be changes of two or more grades. The proportion of grades that were different from the centre assessed grades was smaller than that in the previous studies.

CCEA also recognised that some grades would differ from centre assessed grades by two or more grades. They used that as a trigger for more in depth checking of the results. They also reported on the likely predictive accuracy to compare potential models.

SQA recognised the intrinsic variability in grades. They developed a model based on tolerances and ranges that used the individual information provided by centres. The ranges were wider than historic attainment at the centre, which meant that a centre was not constrained to its historic attainment. They did not set an expectation around how many grades would change or by how much.

Subject experts checked the results of the model. Although this approach did not set quantitative values for measures of quality, it did set qualitative criteria that these subject expert groups were content to verify the results against.

Key quality criteria in all countries were that the grade distributions should be similar to previous years and that attainment gaps between groups should not widen.

Lessons for others developing models

In our view, the qualification regulators had due regard for the level of quality that would be required. However, the public acceptability of large changes from centre assessed grades was not tested, and there were no quality criteria around the scale of these changes being different in different groups.

Others developing models should:

- Be clear on the quality criteria adopted and their acceptability.

3.3.3.2 Quality assurance of input data

The data used in the model are just as important as the model itself, if not more so. If the data entered is of poor quality then the resulting output will be of poor quality, and therefore will not command public confidence, regardless of the model. The importance of checking the data before running a model cannot be underestimated.

Examples from the 2020 grade awarding processes

All countries issued guidance to centres on providing centre assessed data. This included guidance on how to support objectivity and avoid bias in making judgements. All countries ensured that there was a sign off process by the Heads of Centre confirming the accuracy of the information submitted. As an example, SQA produced an online course and a guidance document for teachers and lecturers on producing the estimated grades. The course included a section on bias and ways of overcoming it.

In addition, awarding organisations provided pre-populated data entry portals for submission of judgements. They validated entry data, for example identification of any changes in entry patterns that did not look valid.

There was not sufficient time available for full training on estimating grades and rank orders to be provided to centres, or for moderation of these grades and rankings between schools to be performed. If this had been able to take place, then the requirements of the statistical model in standardising grades could have been reduced. Analysis of the centre assessed grades showed not only an overall increase in grades, but differences between centres in the extent of grade change. For example, one regulator showed us an example of a centre who had previously received the full range of grades for a subject but who had submitted centre assessed grades of A* and A for all students.

Through their conversations with representative schools and colleges, the Independent Review of Qualifications in Wales in Summer 2020 explored the challenges to providing CAGs and rank orders. This included the understanding of the relative uses of the CAGs and rank orders. They noted that Qualifications Wales expected rank ordering of candidates to be the critical data element from centres that fed into the awarding of grades. However, the review team ‘did not gather the same understanding from centres who regarded the CAGs as the key data element and who struggled with a uniform linear ranking of candidates when they knew that, in reality, candidates bunched around performance points’.

The relative use of centre assessed grades and ranking data in the final model was not fully communicated to centres, particularly in England, Wales and Northern Ireland, as this was not known when the data was originally requested.

CCEA and WJEC carried out checks on the centre assessed data - for example, to ascertain that rank orders were plausible in relation to banked evidence. They went back to centres with specific queries.

Ofqual undertook validation checks to ensure that values submitted were valid for each field and consistency checks were made to the data once it was combined across awarding organisations.

SQA reported to us that they had expected the data they received from centres to be better quality. They did not go back to centres when estimates seemed implausibly higher than previous attainment, as there were issues around how they could do this in a fair and equitable way within the time constraints. However, in developing Starting Point Distributions – which defined the acceptable national-level 2020 attainment for each course - subject Heads of Service were involved and used their knowledge, for example, around courses changes, to identify courses where 2020 attainment would reasonably be expected to deviate from historic attainment.

In England, Wales and Northern Ireland where awarding organisations were using common historical data, there was one final version that was quality assured through independent duplication. Measures were taken by Ofqual to ensure consistent approaches were taken by each awarding organisation in using historical and prior attainment data.

Lessons for others developing models

The support and quality criteria that was given to centres and head of centre sign off demonstrated an effort by Regulators and Awarding Bodies to ensure quality within the time constraints. The historical data was also quality assured.

In our opinion, it may not have been clear to all centres how the ranking data would be used in the model and therefore the level of quality that was required in determining them.

The lack of time to provide full training and moderate grades between centres may have impacted on the quality of the input data. In addition, in some countries there was debate around what checking of input data was appropriate and whether it was permissible to go back to centres to query the data that they supplied.

We believe that these factors may have impacted on the quality of the input data.

Others developing models should:

- Be clear with data suppliers and operational staff how the data will be used and the level of quality required.
- Quality assure input data.

3.3.3.3 Quality assurance of the outputs from the model at the level they will be used

The Code sets out that quality assurance arrangements should be proportionate to the nature of the quality issues and the importance of the statistics in serving the public good. Statistics producers should be transparent about the quality assurance approach taken throughout the preparation of the statistics. This includes the use of human verification of results involving individuals or groups of individuals analysing the results of a model to check for and correct unusual model results. This is sometimes also referred to as having a ‘human in the loop’. However, as this term can mean several different things, to avoid confusion we have used the terms ‘human verification’ and ‘human review’ in this section. The risk and impact of quality issues on statistics and data should be minimised to an acceptable level for the intended uses in order to ensure public confidence.

Examples from the 2020 grade awarding processes

Quality assurance checks at subject level were carried out in all countries. In order to verify the results obtained from their models, Scotland, Wales and Northern Ireland all held validation meetings with subject specialists. When they identified issues, they re-ran the models.

In Wales, once the models were executed and grades were produced there was close scrutiny carried out at individual qualification level. This is the normal process every year but this year additional detailed reports were produced by WJEC. These included statistics on how grades varied over time, attainment gaps by gender, extreme difference between centre assessed grades (CAGs) and calculated grades and cross tabulations of CAG and calculated grade for individuals. These reports were checked internally within WJEC and as a consequence models were re-run. The statistics team within Qualifications Wales also examined the reports and raised potential issues with WJEC for review.

CCEA sought additional resource to help with quality assurance including statisticians from the Northern Ireland Statistics and Research Agency (NISRA) and mathematics interns. They contracted an external data science company to replicate testing and live grading. This provided additional quality assurance.

CCEA adopted a range of other quality assurance checks. If individual centre outcomes differed for 2020 compared to three previous years for every grade, these were highlighted and reviewed by CCEA subject specialists. They also identified candidates whose CAGs were two or more grades different from that generated by the CCEA standardisation model. CCEA hired contractors who independently developed code and ran the same data side by side. This ensured that the computer code was 'error free'. They only accepted results when they matched. They also hired mathematics interns to manually check results.

Ofqual told us that there was an acceptance that CAGs would be different to standardised grades, as research literature shows that sometimes teachers' predictions may be significantly out of line with the grade achieved. They also told us that there were extensive validation processes put in place by awarding organisations and by Ofqual overseen by daily contact with the Standards and Technical Issues Group. This included quality assurance activity such as confirming that the model was functioning as intended, reviews at subject, centre and cohort level, identification of discrepancies to allow centres to resubmit where there may have been errors and final checks on outcomes to reconcile and ensure quality assurance processes had been followed. Awarding organisations worked closely together to ensure a consistent approach was taken by each of them.

The regulators and awarding organisations in England, Northern Ireland and Wales met following the running of the standardisation process, to review the overall national outcomes from the awards for both A levels and GCSEs.

Cambridge Assessment, who operate the Oxford, Cambridge and RSA Examinations (OCR) awarding organisation, identified some potential flaws with the methodology used to award grades in 2020. In their evidence to the Education Select Committee they stated 'As the full set of OCR A level results became available internally, Cambridge Assessment began to interrogate the data to understand how the model had behaved, seeking any evidence of issues that might be emerging. On two of these – individual outlier students (high performing pupils in large, typically poor performing centres) and the profile of results for large centres – we had concerns and felt our analysis had reached an evidential threshold requiring us to inform the Department for Education.' They concluded 'that the extent of concern among centres could have been avoided if extensive checking of individual school and college results had been conducted prior to A level results being awarded. Ofqual have subsequently published a report in December 2020, 'Standardisation of grades in general qualifications in summer 2020: outliers Identifying students for whom the standardisation model would be unreliable' which looked at methods to identify outlying students for who the standardisation process could not be relied upon.

Lessons for others developing models

In our view, all the qualification regulators understood the need to quality assure the results of the model and there were clear examples of good quality assurance of output data. For example, the approach of dual coding in two separate organisations provided an extra level of assurance on the accuracy of the code.

However, quality assurance of the model outputs were mostly performed at subject level. There was limited human review of outputs of the models at an individual level prior to results day. Checking for unusual results may help identify errors or anomalies prior to decision making, which should be explored further as part of quality assurance.

Others developing models should:

- Ensure that the analysis of results for quality assurance purposes supports the intended uses of a model.
- Quality assure the outputs from systems including checking that the programme does what it is intended to do.
- Collaborate with other organisations that can assist with quality assurance, for example by dual coding.
- Before a model’s output is used to support decisions, ensure there is an agreed methodology to check and correct unusual results. This should include clarity on the degree of human review required.

3.4 Key Principle – Meet the need and provide public value

Statistical models can command public confidence when the results **meet the need** for which they will be used and **provide public value**. This chapter sets out the components that we have found to be key to meeting the need and providing public value.

3.4.1 Public bodies must ensure results meet the need in the widest sense

3.4.1.1 Engagement with commissioners of the model and others who can help ensure the results meet the need

Engagement with the commissioners of the model is crucial to ensuring the model meets the need. As set out in the Aqua Book ‘The person commissioning analysis must ensure that those doing the analysis understand the context of the question being asked so that they understand the likely risks and can determine what the appropriate analytical and quality assurance response should be. The commissioner must understand the strengths, limitations, inherent uncertainty and the context of the analysis so that the results are interpreted correctly.'

In addition, there may be other sources of external challenge that can help public bodies ensure that the model meets the need.

Examples from the 2020 grade awarding processes

All of the qualification regulators engaged with the policy officials drafting the direction from the minister.

Qualifications Wales provided recommendations and analysis to ministers to say grades should be calculated. Their view was that it was difficult to see when schools would open again, so the option to delay exams was risky. Qualifications Wales were involved in the drafting of the direction from the Minister. Qualifications Wales reported to us that in Wales the minister sets a policy direction that they have to pay due regard to, but it is actually their board that has the legal responsibility of deciding whether to comply and how that will be implemented.

Ofqual were consulted by the Secretary of State in March 2020 on how to manage school and college qualifications in the context of a pandemic. Their advice was that, if possible, exams should go ahead and be held in a socially distanced manner, but that, if there was widespread national or local disruption to exams or the closure of schools and colleges, a method of moderated teacher estimates (i.e. calculated grades) was an option to be considered. Once schools and colleges were closed, the Secretary of State made the decision that exams should be cancelled and issued a direction to Ofqual. This direction set out government policy and asked Ofqual to have regard to it in making the appropriate changes to its regulatory framework. In the context of the closure of schools and colleges necessitated by the pandemic, and the fact that government had confirmed that exams could not go ahead, Ofqual told us that its Board decided it should place significant weight on the Secretary of State's direction in informing the decisions it took.

In Northern Ireland, the Education (NI) Order 1998 sets out CCEA's powers - which are to provide examinations, not qualifications. As they could not conduct examinations in the summer of 2020, they were required and specified to conduct alternative arrangements set out by the Minister of Education. The direction received was explicit as to the arrangements that should be put in place for each examination type differing from other UK jurisdictional approaches.

SQA worked with Scottish Government around contingency planning for the 2020 exams and issued joint statements. In late March 2020 the Deputy First Minister, asked SQA to develop an alternative certification model for 2020.

Once the results were published, ministers in each country offered high level assurances on the robustness of the models but were unable to fully defend them. As a result, grades were re-issued based on the higher of centre assessed grades and the calculated grades.

In addition, the qualification regulators had opportunities for external challenge to their approach through the committees that they report to, but concerns were raised that this was not fully utilised. For example, in their letter to the Secretary of State in November 2020 the Education Select Committee expressed regret “that Ofqual decided not to raise wider concerns about the fairness of the model they were being asked to implement. They had every opportunity to do so when they came before us in June.”

Lessons for others developing models

In our view there was strong collaboration between the qualification regulators and ministers at the start of the process.

It is less clear to us whether there was sufficient engagement with the policy officials to ensure that they fully understood the limitations, impacts, risks and potential unintended consequences of the use of the models prior to results being published.

In addition, we believe that, the qualification regulators could have made greater use of opportunities for independent challenge to the overall approach to ensure it met the need and this may have helped secure public confidence.

Others developing models should:

- Engage with the commissioners of the model throughout and ensure they understand the strengths, limitations and risks of the approach.

- Ensure that opportunities for independent challenge are fully explored when developing an approach.

3.4.1.2 Comparability of models supporting similar decisions

In any statistics production or model building scenario it is important that similar decisions are based on comparable information. If different models are used to make the same decisions about different individuals then this will not command public confidence.

Examples from the 2020 grade awarding processes

There were high levels of collaboration between the qualification regulators and awarding organisations during the development of the models. For example, England, Wales and Northern Ireland share the A-level and GCSE brands. As such they had to ensure that similar standards were maintained. For example, CCEA is bound by statute to maintain a similar approach to other UK jurisdictions. The qualification regulators and awarding organisations met regularly and collaborated in developing the models and approaches to appeals to ensure consistency.

The grades awarded in A-levels and Highers inform similar decisions. These include the allocation of university places. As such, although the models used in Scotland differed from those used in England, Wales and Northern Ireland, the qualification regulators held regular discussions to ensure mutual awareness of each other’s approaches. For example, CCEA reported to us that knew that the maintenance of standards and comparability across jurisdictions would be more challenging if CAGs were awarded as it was evident that the CAGs were very generous compared to previous years' outcomes but that they are also bound by statue to maintain similar approaches to other UK jurisdictions.

Lessons for others developing models

In our view, the qualification regulators collaborated to ensure the comparability of approaches to awarding grades in 2020. This in itself would have improved public confidence in the approaches. However, this did mean that criticisms of one regulator's approach was likely to impact on confidence in the approaches taken by each of the other regulators.

Others developing models should:

- Collaborate with others undertaking a similar task.

3.4.2 Public bodies must ensure models provide public value

3.4.2.1 Acceptability testing of model outcomes with all affected groups

Key to commanding public confidence in a statistical model is ensuring that model outcomes are acceptable to affected groups, particularly those with protected characteristics. The likely outcomes of the model should be tested with the affected groups. Guidance exists on how to test acceptability of data-driven approaches with affected groups.

Examples from the 2020 grade awarding processes

Ofqual put resource into testing the acceptability of their approach. This included segmented focus groups with a range of teachers, lecturers, students and parents from different socio-economic backgrounds to evaluate the public acceptability of their approach. These groups helped them test the level of understanding and acceptance of the arrangements. Ofqual told us that during the early focus groups there appeared to be a public acceptance that using the centre assessed grades without standardisation would not be fair due to the lack of time for effective training and moderation.

Ofqual told us that outcomes of focus groups suggested a shift in public attitudes after the Scottish grades were announced and there was widespread publicity; it appeared that at this point people started to consider more about what the approach might mean for them or their children specifically, rather than the merits of the approach in general terms.

All the qualification regulators considered the public acceptability of not standardising grades. However, some qualification regulators did not appear to be aware of the available guidance on testing the public acceptability of statistical models, such as that in Annex B. As with the guidance around ethical use of data and models, this may have been due to the confusing landscape around relevant guidance.

Lessons for others developing models

In our view, early engagement demonstrated an acceptance by stakeholders that the use of centre assessed grades would not provide fairness to students.

From that point on, where testing was carried out, the focus was primarily on testing the process of calculating grades, and not on the impact of the grades themselves. This, and the limited testing in some countries, may have led to the regulators not fully appreciating the risk that there would be public concern about the awarding of calculated grades.

Others developing models should:

- Use the guidance on assessing public acceptability that is available.

3.4.2.2 Involvement of service delivery staff in reviewing the results

As stated in Part 1 of this report, models used in the public sector usually support decision making by humans, rather than making the decisions automatically. The staff delivering a service know the individuals and their circumstances best. They can provide a useful resource for checking the results of the model prior to decisions being made or being the ones to make the decision based on the results of the model. This human involvement is sometimes termed as having a ‘human in the loop’.

Examples from the 2020 grade awarding processes

In the exams context, the teachers and lecturers know the students best. They would be best placed to identify potential issues with the results and take the results of the model to inform the decision of what grade an individual would get.

There were a range of reasons why none of the countries returned the calculated grades to centres for their verification and input. For example, CCEA reported to us that they had sought legal advice around returning unusual results to schools for verification. The advice was that this could not be done in a consistent way and could result in legal challenge.

During early summer SQA had originally indicated that it would investigate the feasibility of engaging with schools, colleges and/or local authorities to discuss any reasons for the change in estimated attainment. At the Educations and Skills committee hearing on 1st May they stated “in finalising the process, we are looking at whether, as part of the moderation process, we can enter into a professional dialogue with a school if the shape, distribution or volume of attainment at that school looks very different this year—in one direction or another—from how it has looked historically.” However, time and concerns about being able to do this in a way that was fair to all centres prevented them from entering into dialogue with centres in the end.

Lessons for others developing models

In our view, there were clear constraints in the grade awarding scenario around involvement of service delivery staff in quality assurance, or making the decisions based on results from a model. These included concerns around consistency of approach and fairness to some students and time constraints. However, we consider that involvement of staff from centres may have improved public confidence in the outputs.

Others developing models should:

- Include human validation of results, including for individuals, by service delivery staff.

3.4.2.3 Timings and grounds for appeal

In any decision-making process about individuals, those individuals have a right to appeal. The grounds and process for appeal should be transparent.

Examples from the 2020 grade awarding processes

Consultations on the appeals process and decision reports on the outcomes were published by Ofqual, Qualifications Wales and CCEA. The consultations focused on the grounds for appeal, who could make the appeal and the impact of appeals on other students' grades. The feedback they received highlighted concerns by parents and students about individuals not being able to submit an appeal and not being able to appeal the judgements or procedures of centres. As the decisions reports highlight, the regulators did not feel that the awarding organisations would be in the position to evaluate the professional judgements underpinning centre assessment data. All the regulators sought to improve the information provided to centres to minimise issues with the procedures used. Instead, complaints procedures were in place for concerns about bias or discrimination in teacher judgments.

The appeals process was the subject of criticism. For example, the Independent Review of Qualifications in Wales in 2020 identified a number of shortcomings with the appeals process and concluded that there was "a failure to put in place a fair and workable appeals process in 2020 that would deal with the known inabilities of the statistical processes to give a fair outcome to every learner."

In Northern Ireland and England there were also concerns raised about not being able to appeal grades on the basis of the standardisation model. Ofqual told us that allowing appeals on the basis of the standardisation model would have been inconsistent with government policy which directed them to "develop such an appeal process, focused on whether the process used the right data and was correctly applied,"[7]. Appeals were permitted based on where a centre could show that the data used in the model was not appropriately representative.

[7] DIRECTION UNDER S 129(6) OF THE APPRENTICESHIPS, SKILLS, CHILDREN AND LEARNING ACT 2009 DIRECTION UNDER S 129(6) OF THE APPRENTICESHIPS, SKILLS, CHILDREN AND LEARNING ACT 2009, Rt Hon Gavin Williamson CBE MP Secretary of State,

Ofqual, Qualifications Wales and CCEA all planned for a simplified appeals process where the awarding organisations were able to quickly identify and correct any errors that have been made.

In addition, in England an autumn series was provided to allow candidates the opportunity to take exams should they continue to be dissatisfied with their grades.

The appeals process formed part of the Post Certification Review (PCR) in Scotland and included provision for candidates or schools to appeal based on evidence of higher performance. In his report National Qualifications Experience 2020: Rapid Review, Professor Mark Priestley observed that there was more information available on appeals than the models prior to results day. However, he also assessed that the likely impact of the PCR process and its public reception in relation to equity issues could have been better communicated and this would have helped to mitigate the subsequent lack of public acceptability.

All the countries saw the appeals process as a step in the model implementation. However, even with a rapid appeals process, decisions around university places had been made based on the initial grades. While the circumstances of the appeals were different in 2020, it should be noted however that a post results appeal process is part of grade awarding in a normal year.

Lessons for others developing models

The qualification regulators saw the appeals process as an integral part of the approach. However, consultations demonstrated a mixed acceptability of the proposed process.

Based on the evidence provided to us, it is not clear that the appeals process would have effectively dealt with the known limitations of the statistical models in this context.

In our view, the limited public understanding around the appeals process and the perception that students would not be able to appeal their grades reduced public confidence in the results of the models.

Others developing models should:

- Be clear and transparent on how individuals can appeal decisions supported by models based on data from the outset.
- Resolve ‘errors’ and ‘anomalies’ as part of a model and prior to the decision being made.

3.5 Summary

This section has explored the approach to awarding grades in each of the four countries to identify wider lessons for other public bodies looking to develop or work with statistical models and algorithms.

We have found that achieving public confidence is not just about delivering the key technical aspects of a model or the quality of the communication strategy but rather, it arises through considering public confidence as part of an end-to-end process, from deciding to use a statistical model through to deploying it.

We have identified that public confidence in statistical models is supported by the following three principles

- **Be open and trustworthy** – ensuring transparency about the aims of the model and the model itself (including limitations), being open to and acting on feedback and ensuring the use of the model is ethical and legal.

- **Be rigorous and ensure quality throughout** – establishing clear governance and accountability, involving the full range of subject matter and technical experts when developing the model and ensuring the data and outputs of the model are fully quality assured.

- **Meet the need and provide public value** – engaging with commissioners of the model throughout, fully considering whether a model is the right approach, testing acceptability of the model with all affected groups and being clear on the timing and grounds for appeal against decisions supported by the model.

Part 4 – Commanding public confidence in statistical models

This section of the report presents high level findings on the grades awards process which in our view impacted most on public confidence and highlights the key lessons for others looking to develop statistical models to support decisions. These lessons apply to those that develop statistical models, policy makers who commission statistical models and the centre of government. We make recommendations for the centre of government to ensure there is sufficient leadership, guidance and support for those developing statistical models in the future and outline our own commitments to supporting them.

4.1 Findings in the grade awarding context

Against the background of an inherently challenging task, the way the statistical models were designed and communicated was crucial. This demonstrates that the implementation of models is not simply a question of technical design. It is also about the overall organisational approach, including to factors like equality, public communication and quality assurance.

Many of the decisions made supported public confidence, while in some areas different choices could have been made. In our view, the key factors that influenced public confidence were:

The teams in all of the qualification regulators and awarding organisations **acted with honesty and integrity**. All were trying to develop models that would provide students with the most accurate grade and enable them to progress through the education system. This is a vital foundation for public confidence.

Confidence in statistical models in this context - whilst we recognise the unique time and resource constraints in this case, a high level of confidence was placed in the ability of statistical models to predict a single grade for each individual on each course whilst also maintaining national standards and not disadvantaging any groups. In our view the limitations of statistical models, and uncertainty in the results of them, were not fully communicated. More public discussion of these limitations and the mechanisms being used to overcome them, such as the appeals process, may have helped to support public confidence in the results.

Transparency of the model and its limitations – whilst the qualification regulators undertook activities to communicate information about the models to those affected by them and published technical documentation on results day, full details around the methodology to be used were not published in advance. This was due a variety of reasons, including short timescales for model development, a desire not to cause anxiety amongst students and concerns of the impact on the centre assessed grades had the information been released sooner. The need to communicate about the model, whilst also developing it, inevitably made transparency difficult.

Use of external technical challenge in decisions about the models - the qualification regulators drew on expertise within the qualifications and education context and extensive analysis was carried out in order to make decisions about the key concepts in the models. Despite this, there was, in our view, limited professional statistical consensus on the proposed method. The methods were not exposed to the widest possible audience of analytical and subject matter experts, though we acknowledge that time constraints were a limiting factor in this case. A greater range of technical challenge may have supported greater consensus around the models.

Understanding the impact of historical patterns of performance in the underlying data on results – in all four countries the previous history of grades at the centre was a major input to calculating the grades that the students of 2020 received for at least some of their qualifications. The previous history of grades would have included patterns of attainment that are known to differ between groups. There was limited public discussion ahead of the release of results about the likely historical patterns in the underlying data and how they might impact on the results from the model. All the regulators carried out a variety of equality impact analyses on the calculated grades for potentially disadvantaged categories of students at an aggregate level. These analyses were based on the premise that attainment gaps should not widen, and their analyses showed that gaps did not in fact widen. Despite this analytical assurance, 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.

Quality Assurance – in the exam case, there were clear examples of good quality assurance of both input and output data. For input data, centres were provided with detailed guidance on the data they should supply. For output data, the regulators undertook a wide range of analysis, largely at an aggregate level. There was limited human review of outputs of the models at an individual level prior to results day. Instead, the appeal process was expected to address any issues. There was media focus on cases where a student’s grade was significantly different from the teacher prediction. In our view, these concerns were predictable and, whilst we recognise the constraints in this scenario, such cases should be explored as part of quality assurance.

Public engagement – all the qualification regulators undertook a wide range of public engagement activities, particularly at the outset. They deployed their experience in communicating with the public about exams and used a range of communication tools including formal consultations and video explainers, and the volume of public engagement activity was significant. Where acceptability testing was carried out, however, the focus was primarily on testing the process of calculating grades, and not on the impact on individuals. This, and the limited testing in some countries, may have led to the regulators not fully appreciating the risk that there would be public concern about the awarding of calculated grades.

Broader understanding of the exams system: in a normal year, individuals may not get the results they expect. For example, they may perform less well in an exam than anticipated. Statistical evidence and expert judgments support the setting of grade boundaries in a normal year. These may not be well understood in general but, as well-established processes they are able to command public confidence. As a result, when the unfamiliar 2020 approach was presented publicly, people may have assumed that an entirely new, machine-led approach was being introduced, and this may have raised their concerns. This issue of broader understanding would have been very hard for the regulators to address in the time available.

Overall, what is striking is that, while the approaches and models in the four countries had similarities and differences, all four failed to command public confidence.

4.2 Lessons for those developing statistical models

Our review found that achieving public confidence is not just about delivering the key technical aspects of a model or the quality of the communication strategy but rather, it arises through considering public confidence as part of an end-to-end process, from deciding to use a statistical model through to deploying it.

We have identified that public confidence in statistical models is supported by the following three principles:

- **Be open and trustworthy** – ensuring transparency about the aims of the model and the model itself (including limitations), being open to and acting on feedback and ensuring the use of the model is ethical and legal.

- **Be rigorous and ensure quality throughout** – establishing clear governance and accountability, involving the full range of subject matter and technical experts when developing the model and ensuring the data and outputs of the model are fully quality assured.

- **Meet the need and provide public value** – engaging with commissioners of the model throughout, fully considering whether a model is the right approach, testing acceptability of the model with all affected groups and being clear on the timing and grounds for appeal against decisions supported by the model.

Underpinning each principle, we have highlighted learning points which are of relevance to all those using data-driven approaches to support decisions in the public sector.

Key Principle – Be open and trustworthy

Public bodies must be open in the development of models	
Transparency of aims	• Ensure aims are clearly stated, achievable with a statistical model and accepted publicly. • When developing and evaluating a model, regularly assess that the stated aims will be delivered. • Be clear where there are tensions between aims and what the relative priorities are.

Openness to feedback during model development	• Ensure that continuous engagement is sought, and feedback is acted on throughout the model development.
Public bodies must be open in the deployment of models	
Transparency of the model and limitations	• Clearly explain the limitations of algorithms and the approach adopted to all stakeholders. • Make data available to researchers to help evaluate methods. • Be clear if the final model is not known when requesting the input data or engaging with stakeholder groups. • Be clear what aspects of the model have been tested. • Be transparent about the limitations of the available data.
Transparency of the process being replaced by a model	• When using an algorithm in place of another process, communicate the strengths and limitations of the usual process so that all audiences can understand the usual level of uncertainty.
Understanding of the impact of social inequalities in the input data	• Be clear about the social inequalities that exist in the underlying data and how they have been treated.
Public bodies must be trustworthy	
Honesty and integrity	• Be aware of, and if necessary, address any organisational objectives which might impact on the development and communication of a model.
Transparency and communication	• Focus on transparency as well as communication throughout the development of a model. • Ensure that communication strategies support all affected groups to understand the impact of a model.
Ethical use of data and models	• Identify relevant legislation around the ethical use of data and models to ensure that both legal and ethical issues are considered.
Involvement of lead analyst	• Ensure the lead analyst has a role within the governance to support trustworthiness, quality and value. • Ensure that where advanced statistical techniques are being considered the National Statistician should be consulted the for advice and guidance.

Key Principle – Be rigorous and ensure quality throughout

Public bodies must be rigorous in how they develop models	
Involvement of analytical and subject matter experts	• Where there is no national or international guidance or established best practice, it is important to gain professional consensus on the proposed method. • Methods should be exposed to the widest possible professional audience of both analytical and subject matter experts, including the potential for external challenge through Red Teams or independent advisory panels.
Decisions on key concepts within the model	• Be clear how the limitations of the available data impact on decisions on key concepts in a model. • Ensure that a consistent approach is applied to all individuals when making key decisions about a model. • Develop and test a model on the data that will be used to run it. • Assess a model against the groupings that will be affected by the output and those of interest to the public (where these are different).
Public bodies must be rigorous in how they apply models	
Clear governance and accountability	• Ensure the governance and accountability are clear. • Ensure that the limitations of statistical models are clearly communicated to non-statisticians within the governance structure. • Ensure that risks around public acceptability and what is achievable with a model are fully managed within the governance structure.
Public bodies must ensure quality throughout	
Agreed and documented quality criteria	• Be clear on the quality criteria adopted and their acceptability.
Quality assurance of the outputs from the model at the level they will be used	• Be clear with data suppliers and operational staff how data will be used and the level of quality required. • Quality assure input data.

Quality assurance of the outputs from the model at the level they will be used	• Ensure that the analysis of results for quality assurance purposes supports the intended uses of a model. • Quality assure the outputs from systems including checking that the programme does what it is intended to do. • Collaborate with other organisations that can assist with quality assurance, for example by dual coding. • Before a model's output is used to support decisions, ensure there is an agreed methodology to check and correct unusual results. This should include clarity on the degree of human review required.

Key Principle – Meet the need and provide public value

Public bodies must ensure results meet the need in the widest sense	
Engagement with commissioners of the model and others who can help ensure the results meet the need	• Engage with the commissioners of the model throughout and ensure they understand the strengths and limitations of the approach. • Ensure that opportunities for independent challenge are fully explored when developing an approach.
Comparability of models supporting similar decisions	• Collaborate with other undertaking a similar task
Public bodies must ensure models provide public value	
Acceptability testing of model outcomes with all affected groups	• Use the guidance on assessing public acceptability that is available.
Involvement of service delivery staff in reviewing the results	• Include human validation of results, including for individuals, by service delivery staff.
Timings and grounds for appeal	• Be clear and transparent on how individuals can appeal decisions supported by models based on data from the outset. • Resolve 'errors' and 'anomalies' as part of a model and prior to the decision being made.

4.3 Lessons for policy makers who commission statistical models

We have identified lessons for ensuring public confidence for commissioners of statistical models from the perspective of supporting those developing them.

- **A statistical model might not always be the best approach to meet your need.** Commissioners of statistical models and algorithms should be clear what the model aims to achieve and whether the final model meets the intended use, including whether, even if they are “right”, they are publicly acceptable. They should ensure that they understand the likely strengths and limitations of the approach, take on board expert advice and be open to alternative approaches to meeting the need.
- **Statistical models used to support decisions are more than just automated processes.** They are built on a set of assumptions and the data that are available to test them. Commissioners of models should ensure that they understand these assumptions and provide advice on acceptability of the assumptions and key decisions made in model development.
- **The development of a statistical model should be regarded as more than just a technical exercise.** Commissioners of statistical models and algorithms should work with those developing the model throughout the end to end process to ensure that the process is open, rigorous and meets the intended need. This should include building in regular review points to assess whether the model will meet the policy objective.

4.4 Lessons for centre of Government

For statistical models used to support decisions in the public sector to command public confidence, the public bodies developing them need guidance and support to be available, accessible and coherent.

The deployment of models to support decisions on services is a multi-disciplinary endeavour. It cuts across several functions of Government, including the Analysis function (headed by the National Statistician) and the Digital and data function, led by the new Central Digital and Data Office, as well as others including operational delivery and finance. As a result, there is a need for central leadership to ensure consistency of approach.

The Analysis Function aims to improve the analytical capability of the Civil Service and enable policy makers to easily access advice, analysis, research and evidence, using consistent, professional standards. In an environment of increasing use of models, there is an opportunity for the function to demonstrate the role that analysis standards and professional expertise can play in ensuring these models are developed and used appropriately.

Our review has found that there is a fast-emerging community that can provide support and guidance in statistical models, algorithms, AI and machine learning. However, it is not always clear what is relevant and where public bodies can turn for support - the landscape is confusing, particularly for those new to model development and implementation. Although there is an emerging body of practice, there is only limited guidance and practical case studies on public acceptability and transparency of models. More needs to be done to ensure there is sufficient access for public bodies to available, accessible and coherent guidance on developing statistical models

Professional oversight support should be available to provide support to public bodies developing statistical models. This should include a clear place to go for technical expertise and ethics expertise.

4.5 Recommendations

These recommendations focus on the actions that organisations in the centre of Government should take. Those taking forward these recommendations should do so in collaboration with the administrations in Scotland, Wales and Northern Ireland, which have their own centres of expertise in analysis, digital and data activities.

Recommendation 1: The Heads of the Analysis Function and the Digital Function should come together and ensure that they provide consistent, joined-up leadership on the use of models.

Recommendation 2: The cross-government Analysis and Digital functions, supported by the Centre for Data Ethics and Innovation should work together, and in collaboration with others, to create a comprehensive directory of guidance for Government bodies that are deploying these tools.

Recommendation 3: The Analysis Function, Digital Functions and the Centre for Data Ethics and Innovation should develop guidance, in collaboration with others, that supports public bodies that wish to test the public acceptability of their use of models.

Recommendation 4: In line with the Analysis Function's Aqua Book, in any situation where a model is used, accountability should be clear. In particular, the roles of commissioner (typically a Minister) and model developer (typically a multi-disciplinary team of officials) should be clear, and communications between them should also be clear.

Recommendation 5: Any Government body that is developing advanced statistical models with high public value should consult the National Statistician for advice and guidance. Within the Office for National Statistics there are technical and ethical experts that can support public bodies developing statistical models. This includes the Data Science Campus, Methodology Advisory Service, National Statistician's Data Ethics Committee and The Centre for Applied Data Ethics.

We will produce our own guidance in 2021 which sets out in more detail how statistical models should meet the Code of Practice for Statistics. In addition, we will clarify our regulatory role when statistical models and algorithms are used by public bodies.

4.6 Conclusion

What has the review of the 2020 exams process in the UK told us about public confidence in models?

In preparing this report, we have been very conscious of the challenges faced by the qualification regulators. We have sought not to bring too much hindsight to our analysis and findings, but instead to highlight what was done in the four countries, and how different approaches were adopted. We have then used this analysis to identify lessons for others.

We do consider that the qualification regulators could have made different choices, both technically (for example, around quality assurance and use of external expertise) and in terms of public engagement. But the fact that the differing approaches to statistical modelling led to the same overall outcome in the four countries, implies to us that there were inherent challenges in the task; and these challenges that meant that it would have been very difficult to deliver exam grades in a way that commanded complete public confidence in the summer of 2020.

This review of the approach to developing statistical models for awarding 2020 exam results is not about the successes or failures of individual regulators. What this review has shown, is that public confidence matters and, once public confidence starts to unravel, it can be very difficult to recover.

Public bodies should not underestimate the challenge of using statistical models to support decisions about individuals. We have shown through this review that achieving public confidence is not dependant on achieving one or two key technical requirements, rather it is ensured by the multiple decisions and actions taken throughout the whole development and deployment of a model. We have also shown that it can be quickly and unexpectantly undermined by one or two key issues.

Given the complex and potentially confusing landscape, identifying where and who is providing relevant guidance and wider support is a key consideration for public bodies wanting to ensure that their models command public confidence.

Annex A: Limitations of algorithms

This annex provides additional information on the limitations of algorithms and includes examples from the grade awarding process in England.

Algorithms take in data, apply a number of operations, and output a result. Algorithms will perform only as, and exactly as, their internal structure prescribes, but more advanced “machine learning” and “artificial intelligence” algorithms allow adaptive changes to internal parameters and structures in response to changing data. The Ofqual algorithm is of a simple type, not allowing adaptation or automatic updating.

The brief description in the opening sentence of the previous paragraph allows us to characterise what we might legitimately expect from an algorithm, and where our hopes might exceed realistic expectations. Once a potential limitation has been identified then remedial action might be taken, though this may not always be possible.

Part 1 of this note discusses the main sources of limitations of algorithms in general, while Part 2 focuses on the particular issue of irreducible intrinsic variation in examination marks and grades.

A1. Algorithms in general

A1.1 Has the objective been properly formulated? Is it ambiguous and/or does it contain contradictions? Are all the terms and concepts within it clearly operationally defined? To the extent that these questions are not satisfactorily answered, we cannot expect an algorithm to give trustworthy results.

To take a pertinent illustration, in the A-level context it would be unrealistic for the objective to be to build an algorithm to predict “the grade a student would have got had they sat the exam” because, as explained below, there are aspects which are impossible to predict (such as illness on the day of the examination) which will impact the obtained grades. Rather, one might attempt to predict “a grade which reflects their true level of attainment”. The irreducible intrinsic variation in the obtained grades imposes a limit on the accuracy of prediction with which “the grade they would have received had they sat the exam” can be predicted.

A1.2 Have the implications of the algorithm been thought through? In particular, are there constraints imposed which might have unfortunate consequences?

In the Ofqual case for example, were the implications of the requirement that a score distribution should have the same shape as previous score distributions considered?

Constraints can sometimes work in contrary directions. This can impose a limit on what might be achieved by an algorithm, and in a worst case can mean that no algorithm can satisfy all of them. Notions of “fairness” are particularly problematic as there are multiple provably-contradictory definitions. For example, what might be “fair” for a group might be “unfair” for all the individuals within it.

A1.3 Limitations of data quality lead to limitations in what might be expected from an algorithm. Can the algorithm cope with anomalies, unexpected data, or cases significantly different from any previously encountered? Has rigorous testing been carried out, exposing the algorithm to data anomalies, including issues of missing data and errors in the coding of data? To the extent that these questions are not satisfactorily answered, performance of the algorithm will suffer.

In some contexts (e.g. medical diagnosis) a “reject option” can be applied, in which unusual or unclear cases are rejected by the algorithm, for closer expert investigation, for more data to be collected, or for an alternative algorithm to be applied. In other contexts, where the data are relatively uniform and homogeneous and where there is a high premium on uniformity in which the way cases are handled, there may be merit in forcing all cases to be processed by the same algorithm.

A1.4 Many algorithms require parameters to be set, and these are often estimated from previous populations. To the extent that such a previous population is different from and not representative of the population to which the algorithm will be applied, poor performance can result.

A particular challenge arises when previous populations are recognised to be biased in the sense that they disproportionately represent or do not represent certain subpopulations.

A1.5 For predictive algorithms, how is the performance of the algorithm assessed? Has care been taken to ensure that the accuracy measure is tapping relevant aspects of performance? This can be critical, since there are typically multiple ways of measuring performance and “accuracy”, and a high score on one need not imply a high score on another.

In the A-level context, given that “true” level of attainment is unobservable, sophisticated statistical methods will be needed to see how accurately it is predicted (see, for example, Murphy and Davidshofer, 2001; McDonald, 1999; Hand, 2004). The possible value that a performance measure can achieve is necessarily limited by the extent of irreducible intrinsic variation as illustrated below.

A1.6 Algorithms need to be validated. In addition to all the aspects above, code needs to be checked and verified as far as possible. Bugs do occur.

A2. Irreducible intrinsic variation in examination marks and grades

“Irreducible intrinsic variation” means that repeated observations or measurements may lead to different results, and that it is not possible to reduce the extent or range of these differences to zero. The use of statistical experimental design and elaborate estimation procedures sometimes means the size of this variation can be shrunk, but never to zero – it is intrinsic. This variation is sometimes described by the technical term “error” (e.g. in classical test theory, see for example Raykov and Marcoulides, 2011), being variation about a mean value, but it is not error in the familiar sense of departure from a “true” value, so this term has the potential to be misleading outside technical discussions. “Uncertainty” is another term which is sometimes used, but this also has less desirable interpretations, and “variation” best captures the concept.

Intrinsic variation includes:

A2.1 Variability between markers (e.g. the student might have been penalised for a particular style of answer by one marker but not by another)

Exam boards should, and indeed do, carry out careful standardisation procedures to control this source of variability as much as possible. However, if multiple examiners are used, some variation will always remain. One way in which it can be reduced, and a way which has become increasingly used with the shift towards on-screen marking, is to have each question or part of a question marked by different markers. Assignment of questions to markers will be via some formal randomisation procedure. This has the result of diluting away random differences between markers when aggregate scores are produced for each student. However it cannot be eliminated completely.

The extent of this source of variation can be estimated by comparing different markers’ scores of the same scripts.

A2.2 Variability within markers (e.g. lack of consistency if marking spreads over an extended period)

This can be controlled to some extent by careful training of the markers, but some will always remain.

Estimation of the extent of this source of variation requires more subtle statistical estimation procedures, for example, using notions of test-retest reliability from psychometric theory.

A2.3 Variability between occasion (e.g. if the student had been tested on a different day when they were less tired)

This source of variation is outside the control of the exam setters and administrators and so is intrinsic to the final marks and cannot be reduced by the exam setters and administrators. Estimation requires elaborate procedures and methods. Comparison of mark or grade distributions over time measures aggregate statistics (e.g. Ofqual, 2016, which looks at how schools' grade distributions change between one year and another), not how individuals would vary were they to have taken the test on a different occasion.

A2.4 Variability between tests (e.g. if different questions had been asked, for which the student was better prepared, or if the student chose different units within an A-level)

This source of variation is intrinsic to the student, and is inevitable given a range or pool of different questions that might be asked. Estimation is possible but requires careful experimental design.

In converting marks to grades we can add a fifth source.

A2.5 Variability over different teams which might have been chosen to determine the grade cut-off.

Additional contribution by Professor David J. Hand.

References

Hand D.J. (2004) *Measurement Theory and Practice*. Wiley.

McDonald R.P. (1999) *Test Theory: A Unified Treatment*. Lawrence Erlbaum.

Murphy K.R. and Davidshofer C.O. (2001) *Psychological Testing: Principles and Applications.* Prentice-Hall.

Ofqual (2016) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/546607/A_level_centre_variability_2013_to_2016__00000002_.pdf

Ofqual (2020) https://www.gov.uk/government/organisations/ofqual/about/statistics

Raykov T. and Marcoulides G.A. (2011) *Introduction to Psychometric Theory*. Routledge.

Annex B: Examples of published guidance

This annex provides **examples of recent publications** related to the topic areas raised in the learning points of this report to illustrate the range of guidance available.

Organisation	Report name	Published
UK Statistics Authority Office for Statistics Regulation	Code of Practice for Statistics	2018
UK Statistics Authority	National Statistician's Data Ethics Advisory Committee: Ethics self-assessment tool	Latest version online
Department for Digital, Culture, Media and Sport	National Data Strategy	December 2020
HM Treasury	The Aqua Book: guidance on producing quality analysis for government	March 2015
Centre for Data, Ethics and Innovation (CDEI)	Review of Bias into Algorithmic Decision Making	November 2020
Centre for Data, Ethics and Innovation (CDEI)	AI Barometer	June 2020
The Ada Lovelace Institute	Examining the Black Box	April 2020
Information Commissioners Office (ICO) and The Alan Turing Institute	Explaining decisions made with AI	May 2020
Committee on Standards in Public Life	Artificial Intelligence and Public Standards: report	February 2020
Department for Digital, Culture, Media & Sport, Government Digital Service, Department for Business, Energy & Industrial Strategy, and Office for Artificial Intelligence	A guide to using artificial intelligence in the public sector	June 2019

Organisation	Report name	Published
The Ada Lovelace Institute	Ensuring data and AI work for people and society	2019
The Alan Turing Institute	Understanding artificial intelligence ethics and safety	June 2019
European Commission High-Level Expert Group on Artificial Intelligence	Ethics Guidelines for Trustworthy Artificial Intelligence	April 2019
Government Digital Service	Data Ethics Framework	June 2018
Office for Statistics Regulation	Quality Assurance of Administrative Data toolkit	February 2019
Information Commissioner's Office	Guidance on AI and Data Protection	July 2020
Department for Business, Energy and Industrial Strategy	The use of public engagement for technological innovation- Literature review and case studies	January 2021

Annex C: Organisations that took part in this review

Review Expert Oversight Group

Professor Sir Ian Diamond, National Statistician

Professor David Hand, Imperial College London

Sir David Norgrove, Chair of the UK Statistics Authority

Professor Sir David Spiegelhalter, Chair of the Winton Centre at Cambridge University

Qualification Regulators

The Office of Qualifications and Examinations Regulation (Ofqual)

Qualifications Wales

Scottish Qualifications Authority (SQA)

Council for the Curriculum, Examinations and Assessment (CCEA)

Others

Department for Education (DfE)

Welsh Government (WG)

Scottish Government (SG)

Department of Education Northern Ireland (DoE)

Independent Review of arrangement to award general qualifications in Wales

Lead for National Qualifications experience 2020: rapid review in Scotland

Deloitte lead for Department of Education (NI) – Review of Exam Awarding Summer 2020

Ada Lovelace Institute

BCS, The Chartered Institute for IT

Centre for Data Ethics and Innovation

Chair of Inter-departmental Working Group on QA of analytical models

Chair of Ofqual’s External Advisory Group

Information Commissioner’s Office

National Audit Office

National Statisticians Data Ethics Advisory Committee

The British Academy

The Data Science Campus

The Office for National Statistics Methodology Advisory Service

The Royal Society

The Royal Statistical Society

The Universities and Colleges Admissions Service

Validate AI