Centre Judgements: Teaching Staff Interviews, Summer 2020

ofqual

Authors

- Steve Holmes, Darren Churchward, Emma Howard, Ellie Keys, Fiona Leahy, Diana Tonin and Beth Black of the Strategy, Risk and Research Directorate.

With thanks to

- The authors would like to thank the teaching staff who gave their time to speak to us in depth and share their experiences and views.

Contents

Executive summary

The exams and assessments that were due to take place in summer 2020 were cancelled in response to the Covid-19 pandemic. Instead, teaching staff in schools, colleges and training providers in England used their judgement to produce grades (known as 'centre assessment grades' or CAGs) and usually also rank orders of students for general and vocational and technical qualifications (and sub-components/units) from Entry Level to Level 3 which students needed to complete to allow them to progress in their education or to employment.

For most qualifications, institutions needed to provide a centre assessment grade for each student and a rank order of students within the qualification entry at that centre. The CAGs were to represent the professional view of the grade students would have received had assessments been able to take place. Some qualifications only required a grade, while others only required a rank order.

Because of the importance of the centre judgements for student progression, and the extraordinary circumstance that Ofqual (the qualifications regulator in England), the awarding organisations and all the relevant teaching staff faced, we wanted to explore the experience and views of teaching staff that were involved. It is important to us and the wider system that we understand as much as we can about this unusual experience and learn from it. The findings from this study have also been important in helping to shape guidance to teaching staff that will apply for teacher assessed grades in summer 2021.

As well as a larger scale survey (Holmes et al 2021) we also carried out in depth interviews with 54 teaching staff with a variety of roles, qualifications taught and from a variety of centre types, to understand fully their experiences. Our survey and interviews took place in late July and early August, before the A level results day and the subsequent announcement that students would not receive calculated grades – statistically standardised CAGs – but would instead receive the higher of the CAG or the calculated grade. These interviews gave us a great deal of insight into the different approaches taken, and the experiences and opinions of those involved.

The guidance produced by Ofqual and the awarding organisations was the primary reference point for the centre judgement process. From this, centres developed their own approach within the parameters of the guidance. An individual school or college's approach was either centrally-designed by the senior management of the school or college, or was delegated to departments to work out.

The centrally-designed process was intended to ensure that CAGs and rank orders for all qualifications were produced in a broadly consistent way within a centre. Within the centrally-designed approach, there were a spectrum of approaches with a

continuum ranging from strongly data-led processes to processes defined by centre-devised guidance.

Throughout all of our interviews it was clear that data was very important, and because data analysis is a normal part of many centres' routine, it was logical for them to adapt existing data collection and analysis to support the judgements. Sometimes management compiled and shared data files with their departments, along with suggested 'calculations' or ways of combining information to arrive at provisional CAGs and rank orders. There was a variety of flexibility in the degree to which these 'calculations' could be adapted within departments. The guidance-led approach relied more on a detailed plan, specifying how departments should make their judgements, but more of the data analysis and evidence weighting was delegated to departments.

Finally, less frequently in our sample, departments were asked to provide CAGs and rank orders to senior management using their best professional judgement, and in line with the published guidance from Ofqual and the awarding organisations. Departments or individual teaching staff were free to consider evidence and make their judgements in the way they thought most appropriate and accurate. Given that centres were permitted by the published guidance to take such a range of approaches, a centre was not able to raise concerns about its CAGs following the release of results on the basis that another institution took a different approach or that different teachers could have come to a different judgement.

The published Ofqual guidance included reference to taking previous results at the school or college into consideration when making judgements. Centres also knew that the submitted CAGs were to be statistically standardised by awarding organisations for all general qualifications and some vocational and technical qualifications. This meant that historical school or college results were often used to check CAGs. This was done either by centre management, or within departments, sometimes to contextualise the CAGs, or sometimes with a firm direction that the CAGs should not be at all higher. A school or college that took into account the distribution of CAGs compared with grades achieved by the centre's students in previous years will have acted within the guidance.

The evidence used by staff to make their judgements largely consisted of the data files already mentioned, which contained a variety of collected marks, from coursework, class work, class tests, year group tests and mock exams. Predicted or target grades were included, and sometimes also additional data such as reasonable adjustment details or attendance data was used. Within general qualifications with predominantly examined assessment, across all of these varied sources of data mock exams were most important. However, a variety of caveats around mock exams were also noted.

In most interviews the individual attributes of students were also considered and used to make adjustments to the data-led judgements. Student trajectory was frequently mentioned, and different profiles of effort, particularly around last-minute revision, were recognised as being difficult to take into account. Some institutions were very heavily evidence-driven and it was recognised that this last minute effort might be factored in, but judgements had to be based on objective evidence.

While most of these considerations were focused on examinations in general qualifications, for vocational and technical qualifications there were a wide variety of approaches. Generally, there was a lot of evidence available due to the continuous nature of assessment for many of these qualifications and consequently confidence in the judgements made was high. The main difficulties related to the diversity of requirements across vocational and technical qualifications and awarding organisations, and the tight timescales required for some submissions.

Consideration of bias and fairness in the grades was commonly discussed. Schools and colleges took a variety of approaches to minimise any bias, using training, sharing information and data analysis to do so. A lot of consideration of specific groups was also mentioned. Generally, the staff we interviewed were happy that the judgements represented largely fair grades and rank orders, although they recognised that some of the uncertainties above did make some judgements hard, especially around where to place certain students in the rank order.

The methods used to make judgements within qualifications - at department-level where classes had to be combined, or within the school or college before submission - were highly varied. In most cases it was clear that a great deal of discussion had taken place, particularly around the 'hard to predict' students or those near to grade boundaries important to progression. Some institutions wanted to avoid being statistically standardised and so adjusted CAGs to match historical results closely, to the dissatisfaction of some class teachers/tutors. Others allowed some increase in results, citing a tolerance before they thought standardisation would be applied, or specific reasons why they thought their results should be higher this year.

In many cases around grade boundaries, students had received the benefit of the doubt, as most interviewees felt it was not fair to enter a CAG at a lower grade when the student had a reasonable chance of achieving a higher grade.

Many interviewees talked about the pressure they felt. They anticipated a lot of pressure from parents and students when the results were issued, but had largely been shielded from any pressure from this source during the process. However, they felt the weight of responsibility for student futures when making judgements, and some pressure on their professional standing from the media and public opinion.

While there was strong confidence in the validity of the judgements for most students, there was some nervousness about the statistical standardisation process that was to be applied and the fairness for students following this. While the whole

centre judgement process was not perfect it was perceived to have worked fairly well and was often viewed as the best process available in the circumstances.

1 Introduction

In response to the spread of the Covid-19 virus in 2020 two directions were given by the Secretary of State for Education cancelling exams in general qualifications (GCSE, AS and A level; 31st March, Department for Education, 2020a) and exams and other assessments in vocational and technical qualifications (9th April, Department for Education, 2020b). Therefore, in late spring/early summer 2020 staff in schools, colleges and training providers in England were involved in producing centre assessment grades (CAGs) and rank orders for submission to awarding organisations (AOs). These CAGs and rank orders were intended to be the starting point for awarding qualifications to students intending to complete whole or parts of qualifications from Entry Level to Level 3 that summer whose assessments had been cancelled. CAGs were the grades which teaching staff thought students would have been most likely to achieve had exams and assessments gone ahead and the rank orders were the best judgement by the centre of the relative ranking of students within each grade.

Given this unusual situation, and the unprecedented need for staff to make these judgements, we wanted to explore how staff had developed processes to support the production of these grades and rank orders and understand more about the evidence basis used to inform these judgements. We also wanted to understand how staff had responded to and coped in the context of these novel arrangements. These insights will be important for us and the wider system to learn from. Following the cancellation of examinations in 2021 (Department for Education, 2021) the findings contained in this report have also been useful in highlighting factors for inclusion in guidance for teaching staff working on teacher assessed grades.

We therefore carried out two pieces of work with teaching staff in July and August 2020: an online survey and a series of in-depth interviews. The online survey we carried out is described in Holmes et al (2021) and from the respondents to that survey we spoke to a smaller number of individuals in interviews to more fully explore their experiences and views. This report describes those interviews.

It is important to remember that all of these interviews were carried out before the decision was taken to award each student the higher of the calculated, statistically-standardised grade or the original submitted CAGs, in place of their calculated grade. This decision was made on the 17th August 2020. Therefore, our interviewees spoke to us in the context that the final grades would reflect some statistical moderation of the submitted CAGs.

While this detailed, qualitative work was originally undertaken before the August 17th announcement, it is still very much a useful and important contribution to our understanding of the summer in a number of respects and will be useful in the context of 2021. It continues a long standing Ofqual research interest in human

judgement of student performance. The arrangements in 2020 are of interest in their own right for many reasons: understanding what happens when teachers replace examiners in determining students' level of performance, when the judgement is based on a holistic appraisal, rather than as a response to a standardised test, and when the learner is known to the person(s) making the judgement, rather than anonymous. There are most likely many insights and lessons to be learned from this process for any assessment systems which rely upon harnessing teacher judgement for summative assessment.

1.1 Previous research

The production of CAGs and rank orders in England in summer 2020 was an entirely novel situation. While teacher assessment is not unusual in other contexts, and contributes to school-leaving qualifications in some countries, these are not specifically predictions of how students would have performed on assessments had they gone ahead. Rather, they are generally more extended, holistic evaluations of students' performance. In a few instances, something closer to assessment performance predictions have been made. However, the need for centres to grade and also additionally rank order students within qualifications has never, to our knowledge, been required in England, or indeed in any jurisdiction. What follows is therefore a short summary of relevant research, seeing what light previous experiences of making judgements of student performance can throw on the circumstances this summer.

Previous academic literature indicates three key aspects of teacher judgements, which will be explored here. These are related to the evidence used to arrive at their final judgements, the accuracy of these judgements, and the potential for biases to influence the final judgements. The context of these findings are grounded in research looking at predicted grades, such as for university applicants in the UK; and internationally where teacher assessments of achievement are/have been used regularly in place of exams.

1.1.1 How teachers make grade predictions/assessments

Teachers in the UK are routinely required to predict A level grades for their students as part of the university admissions process. These predictions are typically made halfway to two-thirds of the way through the final academic year. Importantly, these refer to estimates of the grades students have the potential to achieve at the end of their studies.

It is found that the main sources of information teachers used to inform their predictions were mock exams, formative class tests, and student attributes (such as commitment, attitude, ability to cope with the stress of exams, and how they responded to formative feedback). Teachers also mentioned that having verbal

discussions with students helped them to gauge students' true understanding of topics (Gill, 2019). In general, the most weight was given to mock exam results, sometimes prioritising these over formative class test results. This was because mocks were perceived as providing the most similar setting to real exams students could experience.

Schools also routinely use prior attainment data from standardised tests to help monitor students' learning and progression through their education, and these pieces of data can be helpful in making predictions for attainment (see Rimfeld et al., 2019). Only a few teachers based their predictions on statistical information (such as ALIS or ALPS predictions), but with the condition that these could be overruled if they conflicted with formative class tests or mock results.

Internationally, teacher assessments of a student's current ability have been used as summative assessments of achievement. For example, in New Zealand teachers made 'Overall Teacher Judgements' (OTJs) about primary-aged (5 to 13 years old) students, which were used as part of the 'National Standards' in mathematics, reading, and writing. When generating their OTJs, teachers in New Zealand considered information from a range of sources. These included in-class observations of students completing tasks, running records of achievement, modelling books, and assessment tools such as standardised tests (Ministry for Education, 2011; Poskitt and Mitchell, 2012). Some teachers also used conversations with students, and relied on their teaching experience to help determine a child's ability.

The authors described a tension between the use of qualitative and quantitative types of evidence, often navigating tacit professional knowledge (or "gut feeling"); objective assessment data; and inter-professional judgement, which was based on a combination of the teacher's tacit knowledge, collaborative discussions between teaching staff, and discussions with the student. Some teachers ensured they had at least three pieces of evidence to triangulate and base OTJs on, but also highlighted the importance of moderation processes to help teachers align with and obtain a deeper understanding of standards.

An assessment process in Switzerland investigating summative teacher assessment reported a similar triangulation process between quantitative and qualitative evidence (Allal, 2013). Teachers of sixth grade students (aged 11 to 12) determined grades that formed the basis for orientating students onto different tracks of secondary school studies (academic, vocational or other). The primary piece of evidence used in this task was classroom tests, but teachers used a method of interpretative synthesis to combine *all* the information they had about a student to make a judgement that was most appropriate to that *individual* student. This included using tests, homework, portfolios of work with reflective comments, and classroom

observations; in addition to discussions with parents, other teachers, and other school professionals (nurses, psychologists).

Other studies have identified similar sources of evidence used by teachers in their judgements of students' current ability. These include objective sources, such as standardised assessment tools, assignments, quizzes, and running records; as well as more qualitative observations of effort, attitude, behaviour and students' contributions to discussions (Cameron, Carroll, Taumoepeau & Schaughency, 2019; Cizek, Fitzgerald & Rachor, 1995). The utility of collaboration and moderation of the grades determined through continuous discussion with other professionals has also been highlighted by many as imperative in building consensus and strengthening the accuracy of teacher assessments, (Allal, 2013; Harlen, 2004; Johnson, 2013; Wyatt-Smith, Klenowski & Gunn, 2010).

1.1.2 Accuracy of teacher predictions/assessments

The accuracy of predicted grades can be ascertained by comparing them to the grades students actually achieved after having sat their exams. In the UK context, findings generally indicate that A level predicted grades were accurate less than half of the time, with approximately 40% of predicted grades matching achieved grades (Gill & Benton, 2015; Skills by UCAS; 2013). Predicted grades were generally more optimistic than pessimistic, with between 41% and 48% being over-predicted, compared with a report of 14% being under-predicted (Everett & Papageorgiou; 2011, Gill & Chang, 2013; Gill & Rushton, 2011; Skills by UCAS; 2013).

The extent of these over-predictions is further highlighted by analyses of UCAS data for all Higher Education (HE) course applicants across the UK (Murphy & Wyness, 2020; UCAS, 2017; Wyness; 2016). These analyses compared predicted with achieved A level points attached to a student's best 3 A level grades. It was found that approximately 16% of applicants achieved their predicted A level grade points, while 75% of applicants' grades were overpredicted, and 8.5% were underpredicted. The majority of predicted grades were, however, within 1-2 points (with 1 point equating to 1 grade).

These inaccuracies in predicted grades, however, are not necessarily a reflection of unreliable teacher judgements. Rather, it is commonly found that teachers use predicted grades as a target to motivate students as opposed to a likely estimate of actual performance (Child & Wilson; 2015, cited in Gill, 2019). Wyness (2016) further found that over-prediction was more prevalent at the lower end of the A level attainment distribution. This may exemplify the intention to motivate lower attaining students, however, it is also possible that this reflects ceiling effects, as higher grades can be overpredicted only to a limited extent.

Turning our focus now to teacher assessments as opposed to predictions, the accuracy of summative teacher assessment can be ascertained by comparing them

with grades on standardised tests. There are mixed findings with regards to the accuracy of teacher assessment. In a meta-analysis of 75 studies comparing teacher judgements of students' academic performance compared to students' actual achievement on standardised tests, Südkamp, Kaiser, and Möller (2012) reported that the accuracy of teacher judgement was fairly high overall (with a mean effect size of .63).

However, other literature reviews have found teacher judgement accuracy to be largely variable across different teachers (Brookhart, 2013; Hoge and Colodarci,1989). It has been suggested that where there are discrepancies between teacher judgement and test scores, this is likely because teacher judgements about students' attainment are combined with judgements about other 'academic enablers' such as their effort and work habits (Brookhart, 2013). This is supported by findings described above concerning teachers including qualitative considerations in their judgements. Harlen (2005) highlights that detailed criteria for teachers would support accuracy and consistency in teacher judgement.

Both Harlen (2005) and Hoge and Coladarci (1989) concluded that while teacher judgements could be accurate, they were not always. In the Swedish context, Lindahl (2007) compared teacher assessments of grade 9 students (age 16) with their performance on national tests and found that overall, teacher assessments were more generous than the test results. Moreover, in line with Wyness' (2016) findings for predicted A levels, it was a common finding that teachers were least accurate in judging low-performing students and most accurate in judging high-performing students (Colodarci, 1986; Hoge and Colodarci, 1989; Begeny, Krouse, Brown, & Mann, 2011). Teacher judgements have also been found to be more accurate for individual test items than for more general ratings or global measures (Gabriele, Joram and Park, 2016; Colodarci, 1986).

Although there is a lack of a better measure of accuracy, an important argument made by reviewers of this literature is that ideally, teacher assessments should not be evaluated based on how well they agree with standardised test scores, or be expected to match them. This is because there are important differences between the two forms of assessment (Brookhart, 2013; Harlen, 2005). Teacher judgements provide a more composite measure of ability, because teachers are able to interact with students and observe all their learning activities on a regular basis (Brookhart, 1994).

1.1.3 Potential for bias

One of the strengths of teacher assessment is that it is multidimensional: it takes into account cognitive and skilled-based aspects of academic ability; as well as the non-cognitive aspects of navigating the social processes of learning, such as behaviour and effort (Bowers, 2011). If these factors are valued they can be incorporated into

the construct that is assessed. This would be a wider-ranging construct than is typically assessed through examinations, for example. It is argued that teachers have a responsibility to judiciously use and evaluate all of the information at hand to make a judgement that is in the best interests of the student (Allal, 2013).

However, this may also be its biggest downfall. By using a range of quantitative and qualitative evidence, centre assessments can be vulnerable to bias. Teacher assessment has long been criticised in this respect, with some rooting the issue in the fact that construct-irrelevant information can influence final judgements (Johnson, 2013; Martinez, Stecher, and Borko 2009; Wyatt-Smith, Klenowski, and Gunn, 2010). A message that has been clear in the literature reviewed here, and one that is held by Kahneman (2013), is that through training, moderation, quality assurance and standardisation, it is possible to reduce unreliability and invalidity of the judgements.

Studies of potential bias in teacher judgements were reviewed as part of an Equalities Impact Assessment by Ofqual (Lee & Walter, 2020), for which a brief summary is provided here. Similar to the distinction made earlier, the main areas of research reviewed in that paper in relation to potential bias were teacher prediction or estimation (e.g. predicted A levels submitted to UCAS), and teacher assessment (compared to exam assessment). In summary, differences between teacher assessment and exam assessment results were sometimes linked to student attributes, but these effects were small and inconsistent across subjects.

In terms of teacher prediction/estimation, the subject type had a small but unsystematic effect; and sex and age had small effects, which were inconsistent across subjects. A small effect of centre type was also observed, though this may be explained by the correlation between centre type and attainment and attainment-dependent prediction accuracy (i.e. the reason predictions from independent and grammar schools were the most accurate may be because students in these centre types achieve higher grades on average, and overall, higher grades were easier to accurately predict than lower grades). Some effects of ethnicity and disadvantage were highlighted, with more over-prediction for some ethnic minority groups and for the more disadvantaged in general. Among high attainers though, there was less over-prediction for the more disadvantaged students. However, it was acknowledged that those effects have not been properly estimated.

1.1.4 The current study

The literature review on how teachers make judgements provides some context in which to explore the way CAGs and rank orders were produced in summer 2020. Our intention of this research is to build a contextualised understanding of the processes undertaken within centres following the guidance issued by Ofqual and the awarding organisations, and the views and experiences held by those involved in

the process. Exploration of these issues was driven by the following research questions:

1. What processes did centres follow to make their judgements?

2. Who was involved in making judgements and what roles did they have in the process?

3. What types of evidence contributed towards the judgements, and how were these pieces of evidence weighted and evaluated?

4. Were issues of bias and fairness considered, and if so how and to what degree were these issues monitored and mitigated?

5. Were there sources of pressure in the process, and how and to what degree were these pressures monitored and mitigated?

6. What perceptions are held around the standardisation process to be undertaken by the awarding organisations, and what are the views regarding how this might impact the CAGs and rank orders centres submitted?

7. Was there confidence in the CAGs and rank orders submitted by centres, and what issues contributed towards these perceptions?

1.2 Differences between the normal assessment process and centre judgements

Before we describe the interviews and findings, it is worth considering some of the main differences in the process and individuals involved in assessment decisions between normal years and the unusual circumstances in summer 2020, and therefore why the whole process is deserving of research focus.

Change in the role of teaching staff

Teaching staff across the country had to switch from their usual position of trying to maximise the performance of their students under the usual assessment conditions, to themselves being the assessor or 'examiner'. In other words, they had to switch from being a formative assessor to the summative assessor.

Anonymity

Centre judgements were made by individuals who knew the students well, making it harder to make a completely impartial judgement. In contrast, examinations are almost entirely marked anonymously, with no knowledge of the student producing the response. Indeed, exam boards have safeguards in place to prevent examiners from encountering examination scripts from known individuals.

Experience and training of assessors

Examiners for normal assessment by examination ordinarily need to have at least 2 years' experience as teachers, then they undergo examiner training provided by awarding organisations, and yearly standardisation – this involves forensic reading of candidate responses and discussion of why features or characteristics of the answers capture certain mark-worthy qualities.

While many teachers are examiners, and many teachers also have a great deal of experience of how students achieve on which to make their judgements, the task of making judgements was carried out by the whole population involved in teaching students, including newly qualified teachers. Without a bank of experience to rely on the CAG and rank order judgements for some individuals might be hard to make.

1.3 A note on terminology

We use the word 'centre' to indicate all the different types of institutions involved in producing centre assessment grades, be they schools, colleges, training providers or other types of educational establishment. Because many of the quotes included in this report use abbreviations or specific terms, we have also included a glossary in Annex A.

2 Methods

2.1 Overview

This research was designed to complement the findings of the teaching staff survey to understand how centres made their CAG and rank order judgements, to gather perceptions from those involved about the process, and to explore issues related to centre assessments for future reflection and consideration. Using these two methods allowed us to explore the issues in breadth and depth, which enabled us to develop sufficient detail and contextualised understanding about the judgements made. The survey provided more quantitative data on the process and experience, while the interviews described here provided the richer detail on individual experiences.

2.2 Interviews

We conducted 54 online interviews with teaching staff involved with the process of making judgements between 28 July 2020 and 12 August 2020. The interviews were carried out via video-call using the software each participant was most comfortable with. The interviews were predominantly one-to-one. Each interview was led by one of six Ofqual researchers, each of whom conducted at least 7 interviews. Some earlier interviews were attended by more than one researcher, with one conducting the interview and the other(s) acting (primarily) as an observer. This was for the purpose of standardising researchers in their use of the interview schedule.

Discussions were guided by a semi-structured interview schedule (Annex B), which aimed to elicit information and views about the process by which the judgements were made. Because different roles had a different involvement in the CAG process, meaning there was differential relevance of some topics and questions, the interview schedule was adapted accordingly. These adaptations are outlined on the interview schedule in Annex B where some questions were for specific roles only.

Although the interview schedule provided a structure and focus for the interviews, the participants or interviewers were not prevented from pursuing useful tangents on anything they felt was pertinent to the discussion. The broad and flexible nature of the interviews meant that we could not discuss every aspect of the centre assessment process consistently. The interviews lasted between approximately 40 minutes and 2 hours (the mean duration was approximately 1 hour). All interviews were audio recorded (with participant consent) and transcribed verbatim by an external transcription organisation.

2.2.1 Participants

Respondents to the online survey were invited to indicate if they would be willing to take part in a follow-up interview and share their contact details. We initially contacted individuals to take part in the interviews with a carefully selected cross-section of roles, centre types and qualifications taught. Partly due to the timing, after the end of term, and at a point in the Covid-19 restrictions when there was some scope for going away on holiday, positive responses were lower than they might otherwise have been. We then invited additional individuals, trying to maintain a representative sample, but we did not have full control over the composition of the final sample given the agreement rates we experienced.

We scheduled 54 interviews in total. Those invited to take part reflected a range of educational contexts based on role, centre type, and subject and qualification specialty. Often, the participants, were involved with making judgements for more than one type of qualification and subject. See the following tables for the number of interviewees across role and centre type (Tables 1), subject (Table 2) and qualification type (Table 3).

Table 1. *The number of interviewees by role and centre type.*

	Sixth form college	Academy	Comprehensive	Further education establishment	independent	selective	Training provider	University technical college	**Total**
deputy head of centre		1	1	2	1				**4**
deputy head of department		1	1						**2**
Head of centre			1					2	**3**
Head of department	3	4	6	3	2	1	1	1	**21**
SENCo			1						**1**
Senior leadership team member		3		1					**4**
teacher/tutor	2	3	6	4	2	1			**18**
Total	**5**	**12**	**16**	**10**	**5**	**2**	**1**	**3**	**54**

Table 2. *Number of interviewees by subject*

Subject	Total
Art and design	2
Business	2
Chemistry	1
Constructing built environment	1
Drama	2
Economics business	1
English	7
Geography	2
Geology	1
Health and social care	3
History	1
Languages	3
Mathematics	8
Music	2
Performing arts	1
Politics	2
Psychology	2
Sciences	7
Sociology	1
Statistics	1
Senior oversight	10

Note. Participants were often involved with more than one subject. The total does therefore not equal the total number of interviews that took place. Some participants held a role in which they were involved with many subjects. This is identified as 'senior oversight'.

Table 3. *Number of interviewees by qualification type*

Qualification type	Total
Applied generals	2
AS/A level	22
BTEC	7
EPQ	2
Functional skills	2
GCSE	28
Other VTQ	1
Senior oversight	10

Note. Participants were often involved with more than one qualification. The total does therefore not equal the total number of interviews that took place. Some participants held a role in which they were involved with many qualifications. This is identified as 'senior oversight'.

Prior to taking part in the interviews, participants had given their consent to be contacted and then they signed informed consent forms before taking part in the interview. They were assured of complete confidentiality and told that any information gathered during the interviews would be used for the purposes of the research only. Participants were able to withdraw from the research at any time.

2.2.2 Analysis approach

The content of the transcripts were analysed thematically by five researchers following the guidance of Braun and Clarke (2006) and Nowell, Norris, White and Moules (2017). This was undertaken using the NVivo qualitative analysis software package There were 3 stages to the analysis:

1 An initial coding scheme was prepared on the basis of the questions that had underpinned the interview schedule. The researchers each used this coding scheme to individually analyse a small sample of the transcripts.

2 Following this initial coding process, the researchers worked collaboratively to add and refine codes on the basis of their analyses to represent emerging themes and information. The researchers jointly developed and agreed upon a final coding scheme.

3 The researchers used the final coding scheme to code all of the transcripts (including those from the stage 1 sample).

At stages 1 and 2 of analysis, the researchers individually analysed a shared transcript to ensure consistency against the final coding scheme, and with each

other. This coding consistency check allowed for further discussion and refinement of the coding scheme. Each researcher was also able to identify if, where and why they were coding inconsistently and could adjust their approach to analyses accordingly.

Throughout the analyses, the researchers sought to validate and refine their understanding of the data by searching for any evidence that exemplified the emerging themes. Passages of text were coded (rather than just individual sentences) to ensure that themes and quotes were captured in context. Researchers had access to both the written transcript and the original audio recording of each interview, allowing them to account for the tone of the discussion wherever there was any ambiguity.

2.3 Interview analysis structure

The findings of the interview analysis explore the views of and processes by which they made their judgements, based on the experiences of those involved in determining and submitting them. It is important to note that findings from qualitative research of this type are limited with regards to generalisability because samples are of insufficient size to be representative of the range and variety of views and experiences, and quoting numbers or percentages could be misleading.

Instead, this approach facilitates a rich understanding of the processes that underpin centre assessment in context. Moreover, the intent of using a qualitative research method was exploratory, rather than to identify the number of individuals with certain views and experiences. Because of this, the qualitative analysis does not primarily aim to quantify how many participants commented on each theme. The survey analyses provides detail at this level.

In our reporting of findings, we present an extensive range of quotes to represent participants' experiences in their own words. These are selected to illustrate themes that emerged from the data rather than to represent the sample or population as a whole. To preserve the anonymity of our participants each quote is attributed to the participant's role and centre type. The term 'interviewee' is generally used when reporting the interview findings, but where a particular process or view was relevant to a specific role(s) or centre type(s), this is indicated.

The findings of the qualitative analyses are set within four broad topic areas, which explore a number of issues within these topics. These are:

1. The centre judgement process
    a. Design of processes used by centres to make judgements
    b. Evidence and processes used by teachers and tutors to make judgements
    c. Final submission agreement and quality assurance process
    d. The Vocational and Technical Qualifications centre judgement process
2. Other considerations in making judgements
    a. Bias and fairness
    b. Pressures experienced throughout the process
3. Beyond the centre judgement process
    a. Standardisation process.
    b. Autumn series and progression
    c. Worries regarding the upcoming school year (2020-21)
4. Overall confidence in the centre judgements
    a. Perceived reliability of centre judgements
    b. Perceived validity of centre judgements

The themes drawn out of discussion within these topics are explored in the chapters that follow.

3 The centre judgement process

This chapter sets out the processes by which centre judgements were made. The discussions highlighted several similarities and differences in approaches, both within and between centres. These were related to three main stages of the process, comprising: the design of the process, the evidence used by teachers to make judgements (and how they used the evidence), and how the judgements were agreed and quality assured within centres. Additional details of the centre assessment process that were specific to vocational and technical qualifications (VTQs) are also explored.

3.1 Design of processes used by centres to make judgements

This section looks at the design of the process for making judgements within centres. This includes the use of guidance and information releases from government and awarding organisations (AOs) to help centres design their process, what their process was and what information was disseminated within the centre. Although the process for internally checking and quality assuring judgements was part of this up-front design, this will be described more fully in section 3.3.2.

3.1.1 Guidance from government and exam boards

The initial planning within centres followed on from the announcements that exams and assessments were cancelled, and centre judgements would be required. These announcements were followed by the issuing of information, advice and guidance from the Department for Education, Ofqual and the various awarding organisations. This guidance was used as the primary information for centres to consider when coming up with their process. In this section we look at views of this information and the guidance issued.

3.1.1.1 Ofqual guidance

Generally, the guidance issued by Ofqual was thought to be useful and clear.

> [The guidance] said all of the things that I hoped it would say, which is: teachers are going to use their judgement, be sensible, use all of the information you've got to make a holistic judgement of where you think the pupils would have got to had they taken the exams in the summer. **deputy head of centre, independent**

> Most of what we went off came from yourselves and actually the guidance that came from Ofqual was really helpful. I mean I don't

> know how many staff read it, but there was enough of us in the department that did read it to make sure it was disseminated appropriately to everyone in the department. And I know certainly I didn't start seeing exam boards releasing stuff until I'd already read a few bits from yourselves. So it was really just confirming it all, giving a nice informatic about other procedures, but by that point I think it was fairly clear how it was going to work. **head of department, academy**

However, the timing of the guidance was not always optimal, often coming at a time that made a prompt response to it difficult.

> The abiding joke that ran through was that if you wanted to find out what you had to do you had to wait 'til seven o'clock on a Friday night before a holiday and then that was when the DfE would publish its latest updates. And so I don't know if it was DfE or Ofqual that published [it], the main info came out on Friday the 3rd of April, so that was the last day of our term, at about three in the afternoon. **deputy head of centre, independent**

One person felt that the guidance around the whole process had been too unconstrained and that they would have preferred tighter control and checks over what centres were doing.

> I think if there had been a set of steps that each school had to, or each department had to show that they had followed, then that may have made a difference. […] If you had to show some sort of objective evidence for it. **teacher/tutor, independent**

3.1.1.2 Awarding organisation guidance

Within the area of general qualifications (GQs) the view was that the information from AOs (also called exam boards in the context of GQs) was also generally helpful. Participants commented on some aspects that were particularly good, such as the use of short video tutorials and summary sheets. Participants noted that, because the process was the same across the exam boards, such materials were relevant and helpful regardless of which exam board they were using.

Some felt the exam boards were quite supportive, for example:

> [Regarding information videos] …actually having the face on screen and them saying this is what we'd advise but if you're struggling get in touch with us. That made me feel a lot better about it, because it made me feel like actually exam boards are supporting us in all of this. **head of department, independent**

While others felt the opposite, particularly around ongoing communication.

> They probably could have done more. They could have assigned people to check in with you, you know, how are you doing it for your English, how are you doing it for your maths, you know, that personalised, everything was well here's the email from [the awarding organisation] you know. **deputy head of centre, comprehensive**

Views around VTQs were less positive. The inconsistency of approaches between AOs, the clarity and timing of the guidance and their communication were generally considered to be problems. Section 3.4 considers VTQs in more detail.

3.1.2 Central or delegated design

Following receipt of the guidance from Ofqual and the awarding organisations, centres had to decide how they were going to respond and how they would make accurate judgements. In most interviews, it was clear that the senior management within centres had taken a lead role in designing the process and disseminating this design throughout the centre with a view to standardising the approach across departments and qualifications. However, there were a substantial minority of interviews where a much more delegated approach was taken, with centres requesting that departments, or sometimes individual teachers/tutors, used their professional judgement in their own way that was in line with the requirements set out in the government or AO guidance.

Within the central design approach there was a second contrast, between a data-led approach where centres set the approach through the use of data files containing compiled student data to be used by departments, and a more guidance-led approach, where detailed centre-devised guidance was sent to departments about how they were expected to make judgements, but the actual evidence used was chosen by departments. In the more data-led approach the data files were usually sent out from the centre management (or the data handling side of it) to departments. In the more guidance-led approach departments often had access to their own data sources and compiled these in line with the centre plan. There was a continuum of approaches though, as there were many cases of centres sending out detailed data files but leaving the weighting of different pieces of evidence for the departments to decide.

In all cases, following the determination of the initial judgements for each qualification, there was some quality assuring of the judgements, sometimes within departments but always involving checks by management. These quality assurance (QA) processes will be described in section 3.3.2.

3.1.2.1 Central data-led design

Many centres held detailed information about students centrally, including all of their test and assignment data, and information on attendance. In these cases, a centrally driven process was implemented with both internal guidance and a data spreadsheet that was sent out to all departments, usually to be filled in with grades/rank orders. This spreadsheet was sometimes already filled in with provisional CAGs for the departments to consider as a starting point.

> So the deputy head distilled it down into documentation, which then established guidelines for subject leaders. […] So every single subject leader [had] the last two years' worth of report tracking data that had been put in. Subjects like myself where we had done mocks in January and we'd also done mocks in March, I added that data in as well. So that was the starting point. **head of department, academy**

> [Management] produced a spreadsheet with each of those [evidence categories] and with an explanation of what grades we were expected to input onto that system. Together with, they managed a spreadsheet with previous major examinations, so year 12 examinations, mock examinations and things like that. So it was pretty well set out on the spreadsheet. **teacher/tutor, academy**

> Initially the head teacher gave guidance on how he wanted us to go about producing the teacher assessment grades […] And then so once we had an overview of what the structure was to be, in terms of the whole process, I started by putting together the tracking document. So just making sure that all information in there was available so that we could send it out to the department. That included their previous mock scores that they'd done and their assessment data. **senior leadership team member, academy**

As noted above, this data summary was sometimes sent out as a starting point for departments to modify and apply their own weighting for different types of evidence to.

> We provided the heads of department with a […] template data table where we said look, here are their trial exam results, here are their grades data that we've had through the year so far, here's their past performances, you add in anything else that you think is relevant, cross-set tests, end of term assessments, whatever it might be. And the head of department chose the weighting to give to each particular thing. **deputy head of centre, independent**

> One of our senior leadership team does a lot of stuff with feedback and engagement, so she was absolutely on it straight away. She's a chemist so it's always like spreadsheets. So very quickly she had put out a how-to video as to actually do all the grades and so we [had] very clear guidance which was really useful. […] We'd added a few little tweaks of our own to make it work for our data. So it was very clear. **teacher/tutor, sixth form college**

3.1.2.2 Central guidance-led design

The most common approach mentioned across all of the interviews was for senior managers in centres to share detailed guidance, principles or a plan, often using the Ofqual/AO guidance as a starting point but then with additional detail added, from which departments devised their own approach. This guidance usually specified the general process to be used to make judgements but left departments with the freedom to evaluate their own evidence base and decide how to use this. This central guidance imposed some degree of standardisation across departments. It often promoted the same approach of data files listing student marks, but in this case they were usually devised semi-independently by the departments themselves. This guidance-led approach was often particularly detailed in its quality assurance processes. This stage of the process is detailed in section 3.3.2.

> I was certainly involved in that decision, but it came from management and wider than that it came from our academy trust, because they tried to use the same approach in each of the schools. So yeah, it was really a case of I came up with an idea, made sure it agreed with whatever management wanted and we went from there really. […] I just tried to make sure that the approach was the same in all three sciences. So I spoke to the head of biology, head of chemistry, head of physics, and just made sure that they were happy with the weightings I'd put on. And we had agreement that actually within the A-level course in science this is how we're going to generate the grades. It was then up to the head of subjects if they wanted to tweak that to make it fit their data or not. **head of department, academy**

> The deputy head spoke with the subject leaders and coordinators and laid out what the process would be and we started off by, you know, we had to come up and submit a list of objective criteria that we would be judging the CAGs based on. And so obviously for each subject that's different. **deputy head of department, comprehensive**

> Our starting point was to ask the lecturer to draw up a rubric of evidence in conjunction with the head of department. So they had

> meetings with head of departments and peers where they draw up a rubric of what they considered to be good qualities of evidence. And that was broadly based on exam board specs. **deputy head of centre, further education establishment**

Department decisions and the plan for the data or evidence they would use were often checked and approved by management before the judgement process started.

> We were given a spreadsheet first of all and we had to create what was called a department mark for the pupils. […] We could choose which, what data we wanted to use to create this mark. We were asked quite specifically to try to avoid any kind of subjective mark. So it wasn't just what the teacher thought, it was an actual physical grade or mark that we had that we used. […] Before we were [entering] any of the marks we had to describe to the senior leadership what our, how we were going to generate our mark. So that they could tell, they could cross reference whether it was subjective or not. **head of department, comprehensive**

> Our deputy head teacher had come up with a process that they were going to be using in his subject which is music and talked us through the approach that they were using there. We did have a bit of permissiveness about which model we would use, so what data we would include for each of our subjects. So we weren't given a particular format because some subjects have non-examined assessments that they'd already done coursework for. My course doesn't have any coursework in it at GCSE as such that's assessed. So yeah we were given the opportunity to come up with our own [approach]. And we were allocated a member of the senior leadership team to work on that with. **head of department, comprehensive**

It is worth noting that for further education establishments who wanted to have central oversight of departmental process, the diversity of approaches across VTQ AOs did create a lot of work for senior leaders, and a lot of meetings.

> Because we had that huge variety of awarding bodies and cohort sizes, we felt that we needed to have a process. So [a colleague] and I decided the process. I then briefed the heads of area and took questions. We had one to ones [with] them and we worked through some of the complexities for their area and their individual students. And then pretty much from that we communicated very clearly to the wider college management team as well what we were doing. And from that the managers went and worked with their teams. You know, the number of meetings that they had to have is just incredible to be able to get to the level of detail that we knew we needed at that point. **deputy head of centre, further education establishment**

3.1.2.3 Teacher/tutor or department designed process

As described above, in a substantial minority of the interviews centre management had asked their staff to use their own professional judgement to determine CAGs and rank orders. In these cases management did not specify an approach and didn't usually provide any detailed guidance beyond that provided by Ofqual or the AOs. Staff used judgement to compile and analyse their own data, and their own knowledge of the students to decide CAGs and rank orders. This was often coordinated at a departmental-level by the head of department, but sometimes the teacher/tutor worked alone.

> We didn't get much input. We just got told […] that we had to make a decision based on […] what we knew about them academically obviously, where we'd seen it. So we had to be able to back up our decisions and to go away and do it really. **head of department, selective**

> Teachers were encouraged to reflect on [previous years' data] as they were then planning. I think it varied from teacher to teacher. I think some teachers will have looked long and hard, I think others will have pretty much ignored it and gone no I think that student will get that, and just predicted the grades. **head of department, sixth form college**

> So the school did an audit […] And they said that they weren't expecting departments to all do it the same way, they recognised that some departments had a heavier amount of coursework, or some no coursework, some had marked all their coursework, some hadn't. Whilst we'd all done equivalent mocks at the same time, because they are fixed, they also understood that some people might know for certain students that there was circumstances that affected the mock grades, either all of their mocks or just within that particular subject, where maybe they had catastrophically answered the wrong questions, or there'd just been something. So we were given a lot of autonomy. **head of department, comprehensive**

A number of interviewees reflected that little instruction or guidance was actually needed since the judgement process for the grades was so similar to their normal estimation/prediction practice. Therefore, this less tightly controlled design was felt to be completely sufficient in many cases.

> They just told us what they needed to be honest. In effect we already produce centre-assessed grades with all the estimates we produce; that's not that dissimilar to centre-assessed grades anyway. So there wasn't really any training to do because we kind of did it anyway. **teacher/tutor, comprehensive**

> I was told by my head of faculty what the expectation was. And to be honest even at that point the instruction was it's what we normally do, however we need to also put them in rank. And there was very little more to it than that to be perfectly honest. **teacher/tutor, comprehensive**

Some heads of departments noted how they didn't think a standardised approach across qualifications would have been sensible or manageable due to the differences in the structure of the assessment.

> No I think we pretty much decided that ourselves. I think, because of the amount of coursework we have as well, so I think it's quite hard to have a whole school approach because of the amount of coursework that we, because there's 60% coursework for music compared to none for some of the subjects, so I think it was harder to do that. **head of department, comprehensive**

> I think they [management] gave all that they could, but I don't think there was a totally consistent approach. But then at the same time I don't know how they could have made it consistent, because in a sense it's too late. Because for example some teachers do loads of marking all year, others do hardly any. So they can't really make it consistent, because some people have lots of data, and some don't have much data at all. So yeah, because we didn't know it was coming, how do you make it consistent? **head of department, further education establishment**

Generally, the way judgements were made for VTQs was left to departments due to the diversity of VTQ qualification structures and AO requirements.

> We've got 24 different awarding organisations excluding HEIs. Each of those 24 different awarding organisations had a different view on how and what they were going to use as a calculated result and how we were going to do that. The exams team had extreme pressure to manage each AO's different system and the thresholds of evidence required. **deputy head of centre, further education establishment**

3.1.2.4 Use of previous years' data as part of the centre judgement process

As well as using data about the current students for whom judgements were required, previous years' data for the centre and their departments were almost always used as part of the process to try to ensure that the distribution of the CAGs were the same as, or not much different from, the results in previous years. The Ofqual guidance had mentioned the consideration of previous results as part of the judgement process. In addition, centres were aware of the way the statistical standardisation was expected to operate, by taking previous years' performance into account. This historical data was either used during the production of the CAGs within departments, or it was used to check initial sets of CAGs during a later QA process.

When shared at the start of the judgement process, this information could be sent out by senior management.

> I sat down with my […] data deputy, and we looked at historic[al] performance at a departmental level for each grade, to the extent it existed, and I was very relieved to hear the standardisation […] process matched what we did, which is good. So we looked at two or three years of past data, we took averages. We looked at what kind of progress do pupils tend to make between mock exams, how reliable are predicted results, and so on, and we came up with a set of data. And effectively, so we then came up with a set of grades that we thought each department should be likely to get percentagewise, given the cohorts that they've got. **deputy head of centre, independent**

> One of the senior members who oversees data had sent out to us what, basically a suggestion as to what would be a reasonable value added score based on the last three years, based on what last year was as well, and a bit of an idea of, almost I guess a bit of an alarm bell that if the grades are coming out way above this number that's probably going to look at bit odd, because it's very different to what they've done previously. **head of department, sixth form college**

Often, departments themselves held this data and as part of the process carried out the comparison of the CAGs and previous years' data themselves.

> The school did give a directive and […] they advised that the grades should be the average of your last three years grades. […] So we took a distribution curve the last three years and we tried to work out what our average distribution curve would have been, if we took the

> average of those three years, what would we have got this year?
> **teacher/tutor, independent**

Sometimes an expected grade profile was used up-front before teachers/tutors started working on their class CAGs. These teachers/tutors were given a target to achieve, which varied in the degree to which it was compulsory to meet. Some reported a 'hard target' whereby pre-determined numbers of students had to be allocated grades.

> And [management] also put together the data from the last three years. So what percent got As, Bs, Cs, Ds, and they said that we should make sure that our predictions are in line with the previous data. **head of department, further education establishment**

Some reported a more flexible target, whereby the numbers of students allocated to each grade was negotiable, providing they could support their grades with evidence. One centre calculated these allocations, then reduced them in order to account for the generosity (or benefit of the doubt, see section 3.2.8) that they expected to see in department CAGs.

> We looked at two or three years of past data, we took averages. We looked at what kind of progress do pupils tend to make between mock exams, how reliable are predicted results, and so on, and we came up with a set of data. […] So we then came up with a set of grades that we thought each department should be likely to get percentagewise, given the cohorts that they've got. […] We then reduced those numbers. So we took down by 15% the number of pupils achieving each of those grades. So if there had been 100 pupils getting an A at A-level maths we put in 85 to get an A at A level maths. And then we gave those numbers to the heads of department and said here's your starting point, use these numbers and we wouldn't expect it to be worse than this but use this as your starting point. Because we knew that they would have lots of cases where they'd want to be generous, they'd want to err on the side of positivity, and that then gave us the flexibility to build back in so that we'd come up with a set of results that weren't ridiculous effectively. **deputy head of centre, independent**

Some centres asked teachers/tutors to only produce a student rank order, and then these were allocated to grades based on a pre-calculated expected distribution of grades.

> We got our rank order eventually to a point that we were happy with before we put any grades on. And then […] the top three would get As, the next five would get Bs or whatever it was, and then we looked

> at what our suggested value added would be, which I think was minus 0.35 […]. So we basically went through and went right, to be in line with the previous years, […] from this mapping of grades that would give us a value added of zero. **head of department, sixth form college**

> Once we'd got a reasonable ranking list, we then allocated grades based on what we knew our grades were the last three years since the syllabus started. […] And we worked out we could have three A*s out of 20, […] we could have eight As. We could have six Bs […] or whatever. **teacher/tutor, independent**

A more frequent approach was for more senior management to take the initial sets of CAGs and rank orders from departments and use previous years' performance data to adjust them. This was done either with the knowledge of the teacher/tutors, or behind the scenes without the involvement of the teacher/tutor. The internal moderation of centre CAGs including by using historical data, is considered in more detail in section 3.3.2.

3.1.3 Discussion and training before making judgements

During and after deciding the design of the process to use there were often external sources of information or discussions externally about how other centres were going to approach making judgements. Within centres, discussions and training also took place to inform staff about the process.

3.1.3.1 Support and discussion across centres

Centres often collaborated on the process, either formally, or just through personal contacts. This was often useful in refining ideas, or just providing reassurance to staff that what they were doing was sensible. Sometimes support came through official channels or links between senior staff in centres.

> Like many head teachers, I'm in ASCL, and ASCL were generating quite a lot of guidance. Also, specialist schools and academy trusts were running lots of online forums. And I would always say to staff, 'look, I've sat through about 200 hours of fora about these things and that's what's informing me'. […] So, there was support through the Baker Dearing Trust and we kind of agreed what we would do as a set of UTCs. I mean I wouldn't say that we agreed every single part of the process very strictly. Obviously, we had to adapt that to our own circumstances. **head of centre, university technical college**

> We liaised with a number of other schools. So I had a version of a guidance document from [another school] who are excellent, and my deputy's husband worked at [a second other school], so they had a policy I could look at as well, and we were just sharing. I think lots of schools probably just shared general approaches to say, you know, 'how are you going to go about this practically?' **deputy head of centre, independent**

But there were also a great deal of informal links between teachers discussing, and to some extent aligning, their different approaches.

> My wife's also a maths teacher but in a different school, so she's actually got some TLR [Teaching and Learning Responsibility] within her school. […] So as a couple we did discuss what we were doing. I know that she did something in a similar way to what I'd done, and I can't remember who initiated that idea. But we discussed the pros and cons, and how we might work that, and get a fair way of organising the students. So there was some discussion from a different centre, I guess [a different] point of view. **senior leadership team member, academy**

> So for A-levels for instance, […] sometimes we've only got one teacher for a subject, that teacher was pretty much working on their own. But what they did, which was really fascinating, what they did was they reached out to their other colleagues in other organisations through their networks externally. […] I was quite impressed by that, I thought that was really best practice when they were working in isolation, they were still able to have some support from like-minded colleagues with that expertise. **deputy head of centre, further education establishment**

3.1.3.2 Internal training and discussions

Before the process of making judgements began, staff were often given some training, or at least provided guidance documents or summaries. Often the guidance from Ofqual and AOs was incorporated in a more condensed form in internal training documents.

> Lots of the time I would break down the information to an easy read, which is what we did, myself and another one of our academies within our trust, and then we would share a link to the full document for the staff. Because some of them were quite lengthy documents and how many staff are going to sit down and read that and understand all of that. So it needed to have an easy guide and then here's the link to

> the full document if you want to read it. **deputy head of centre, comprehensive**

> I provided guidance for the teachers, because obviously there's a lot of information that was coming through, but also I shared. So as information came through I would speak to the teachers and say look I'm sending stuff to you, but I will summarise it for you. Because obviously this was, there is no point in everybody reading absolutely everything, but it was there for them to access if they wanted to. **head of centre, university technical college**

Sometimes explicit training on how to make the judgements was developed by the centre. Section 4.1 on considerations around bias gives more detail about staff views on bias training and bias in general.

3.2 Evidence and processes used by teachers and tutors to make judgements

Following the decision within the centre on how the judgements should be made, teachers then undertook the task of assigning grades and rank orders to students using the available evidence. As we saw in the previous section, the design of the process could be roughly categorised into 3 types, a centrally designed data-led approach, a centrally designed guidance-led approach and individual department or teacher/tutor approach. Within each of these designs a variety of different sources of evidence could be used, and we cover the main types of evidence in this section.

This section focuses on making judgements for individual classes. The discussions that occurred between teachers and within departments to co-ordinate and agree the final submitted grades and rank orders across classes are covered in section 3.3.

3.2.1 Data files to capture and weight all available evidence

Across all 3 design approaches, the use of data files representing all the different pieces of assessment and performance evidence was very common. This was usually the starting point for making judgements, as it brought together a lot of information in one place. Even in teacher/tutor-designed cases, they often compiled all their available data. The only difference was that the contents of the data files were less variable across qualifications within a centre in a centrally-designed data-led approach.

It is probably not a great surprise that data was so heavily relied upon, since the requirement to produce CAGs was arguably only a slight extension of normal practice in many centres. A large number of centres today have strong data capture

and analysis processes as a matter of routine, tracking student progress and predicting likely outcomes, to help them continually look at ways to improve their teaching and results. This on-going monitoring process often includes regular class- or whole year-group tests, that feed in to constantly updated predicted grades. These predictions are often reviewed against actual qualification grades each year in order to refine the process and improve prediction accuracy. Therefore, it would have been a simple matter for centres to adapt their existing systems to support the judgement process. This was touched on in section 3.1.2.3, and section 6.1 further looks at why this also gave staff and centres confidence in the judgements they made.

This data-led approach maximises the use of objective data, such as marked class work, class tests, mock results and coursework marks. These were usually presented in the form of spreadsheets, sometimes with a weighting system applied to the marks to combine the evidence, designed to more heavily weight the highest quality and most recent data.

> [We had a] spreadsheet […that] had current working grade[s] based on the two years of assessments, any mocks, and things that were in there and sort of, we do end of term, end of year tests. We don't call them mocks but they are. So all of that sort of stuff. Coursework, […] that was something that I did take into account up to a point as well. So it was a very good way of actually drawing in lots of different bits of information. So the spreadsheet had average GCSE grades, attainment-8 scores, all of those sort of things in as well. So we had a really good range of information to use. **teacher/tutor, sixth form college**

> To generate a data-esque grade my head of department put in a weighting system that we tried to make as fair as possible, which put most weight on the year 11 mock exams they had done, because they're the ones they're most likely to have worked for. But it also factored in the old assessments they had done. And for the most part this gave out a fairly reasonable grade for most. **head of department, academy**

In a few cases the use of data files was seen as the fairest way to determine the grades, reducing the influence of personal biases or feelings about students. This will be picked up in more detail in section 4.1.

> By going data heavy we feel protected and stuff. […] You're working with an individual and a personality, it's that wider interaction with them, […] and that's why we felt it was safest to use the raw data. **teacher/tutor, independent**

> Because you didn't know how the rank order was going to come out and there were a couple where you thought 'oh, so and so has come out higher than I thought'. And then we thought well that's right, we've taken the necessary […] overall impression. **head of department, independent**

Some centres anonymised the process as much as possible in order to minimise the potential for their personal feelings about students to introduce unconscious bias.

> We talked about student numbers and student initials, we didn't talk about individuals. So […] even if staff were presiding over a decision, they would not be talking about the student's name or their number, they would be talking about fact: I've got this evidence, the student has this profile. **deputy head of centre, further education establishment**

The weight given to more qualitative considerations varied when a data-driven approach was taken. Although in many cases the spreadsheets contained just marks from assessments, sometimes they included entries for factors requiring a more subjective evaluation of their effect, such as reasonable adjustments, attendance, or just the teacher/tutor's adjustment factor for each student, with appropriate weighting.

> I set up a sheet which combined any assessments they had throughout the course, so where at A level it incorporated the year 12 and year 13, any internal assessments that had been marked by a teacher. Had a bit of extra context in there, like homework performance and stuff, just to give teachers that extra context of seeing the continued performance and potentially the trajectory that those kids were on across the course. **head of department, academy**

> I put together a spreadsheet of all the assessment marks that the students had had throughout the year. […] The way to make it the least biased was to literally award a score for each assessment, their work ethic, their attendance, and then they all came up with a total at the end, and then that helped us to rank them. **head of department, further education establishment**

> My manager sent a spreadsheet out which broke it down into the different assessments and then she also asked us to think about whether we were basing it on the assessment, to think about the

> progress made as well when we were predicting the grades, which was fair enough I think, but very difficult also. […] But she also, on that spreadsheet that she sent there were other things to take into account, like special characteristics, learning disabilities, so that was the process and then I filled that in formally. **teacher/tutor, further education establishment**

In central guidance-led approaches departments often decided the allowable evidence they would include and then submitted the data plan to senior management to be agreed.

> We had to […] submit a list of objective criteria that we would be judging the CAGs based on. And so obviously for each subject that's different. For example in English we've got no controlled assessments whereas other subjects might be able to use that. So myself and the head of department put that list together, […] obviously you've got mock exams down to maybe classroom. We did have a bit of debate about kind of quality of classwork and whether that could be objective. But, you know, it allowed us to cover some grey areas where maybe students who'd had a bad day in the mock exam but you've seen real evidence in their class work that they'd taken on board feedback and made improvements. **deputy head of department, comprehensive**

Although this data-led approach was by far the most common approach in those we spoke to, it was not universally adopted. Several interviewees discussed more holistic approaches. Often, this started with a set of grades or rank orders that came mostly from their knowledge of the students and their work in a holistic sense, rather than through calculation. This could then often be checked or moderated by data.

> Stage one was 'give a grade and then rank the class'. **SENCo, comprehensive**

> Initially […] we asked the teachers […] to submit what they thought were their predicted grades. And they had a lot of data to go on. And those submissions were sent to me, just for me to give some initial feedback on. […] It was just the evidence that they had, previous assessments that they had, exercise books, just a little bundle of material for those students. And then of course we've got our own data. […] they have already had predicted grades. **head of centre, university technical college**

If there was any pattern to the approaches used, it was that in larger centres, or in departments with larger cohorts, the data-led approach dominated, while in smaller

centres or departments, more emphasis was placed on knowledge of the individual students.

> And I think that's possibly easier for us as a smaller school, because we knew those personalities, but we did try to take that into account. **head of centre, university technical college**

There was also a slight tendency detected in the responses that the more mathematical and scientific subject areas were more data driven, with more creative subjects less so. Although this finding is tentative, based on only a few observations, this may reflect the tendencies and skills of the staff in the respective departments, or the nature of the judgements they had to make.

Although use of data files appeared to be the dominant approach in our sample, it is important to note that in most cases, the judgements did include adjustments based on more qualitative considerations, using knowledge of the students as individuals. It was only in a minority of cases that a solely formula-driven approach was adopted. Section 3.2.4 describes the way more qualitative considerations were used.

3.2.2 Mock exams

Across all of the interviews, the most frequently mentioned, and important source of evidence for making judgements, particularly within GQ, were mock exams. This was true whether a data spreadsheet was being used with weightings applied to different evidence, or a less formal approach. For some, the mocks provided almost all the information required to decide CAGs and rank order and they believed that mocks provided a very close mapping to final qualification grades each year.

> We had just done our paper 2 mock in March […] when the school closed. So we actually had evidence for where the students were at on the second paper. […] So we had two very powerful pieces of data, which I would say drove the grade setting for about 85% of the students. Because the data we felt was representative and the teachers were saying yeah this is a student who's been on a steady flight path and everything's normal. **head of centre, university technical college**

> So for example like the mock data, we gave more weighting to [that than] class assessments just based on the way in which they're sat, obviously there's more rigor in the process of a mock isn't there, than there is in an end of unit assessment. And also factoring in the idea that it's enabling them to focus on revising the whole content of the GCSE rather than just a specific narrow aspect of it. And equally it gives them the opportunity to show what they can do when they're pulling it altogether. **senior leadership team member, academy**

> I would say the school I'm in currently, it seems to me that what students achieve in their mocks is typically what they get. **teacher/tutor, comprehensive**

However, some interviewees did raise issues with the mock exams being used as a strong source of evidence, citing concerns around them not adequately reflecting the students' ability due to, for example, variations in effort (this is considered in more detail in section 6.2). There was also some evidence that there were subject differences in the reliance on mocks, with mocks being a better predictor of final grades in some subjects than others.

> One of the interesting things we found was that some subjects really just wanted to use their mock exam data. Which is fascinating. […] 'We'll use that to rank order the students. So my 200 students for maths GCSE, we'll use the most recent mock, put them into order, put some grades in and we'll use that as a starting for discussion'; whereas other subjects flipped it around and did it the other way and were much more holistic, so things like the arts and humanities, which of course by the very nature of extended writing and extended performance had a difference nuance to them. **senior leadership team member, academy**

Others cited the lack of standardisation in the mocks, with variable teacher marking being mentioned in several interviews, either through no internal process of standardising marking, and/or through the varying ability of the teachers to mark accurately.

> The grades that were coming back from the mocks, in particular for one of the sets, seemed to be quite inflated based on what I know we can achieve as a school. So I was left thinking 'well, have those two teachers quietly done an amazing job this year and if so what's gone on', or is it an issue with the marking? So then we had to go in and start, re-look at some of the marking on the mock papers, which we don't normally, because it's not crucial, we don't normally put all of that effort into standardising the mock marking. **head of centre, university technical college**

In some centres, different papers had been sat as mocks by different classes, making the comparison and combination of students into one rank order difficult:

> We don't mandate the assessment that they undertake formatively. […] One teacher is choosing to use paper X because they feel that that suits the needs of their cohort, but the other teacher is choosing

> to use past paper Y because they feel that that better suits the gaps in knowledge that their cohort has. **deputy head of centre, further education establishment**

Usually these kinds of anomalies in mock results could be spotted in the data-led approach where they stood out from the other assessment data. In these cases their influence could be mitigated (to some extent) by looking at the other assessments or considering teacher judgement.

> And ultimately any complete anomalies were thrown up through the spreadsheet that we used. And an anomaly could be thrown up and if it was quite right and there was justification for it, there were reasons for it. **head of centre, university technical college**

> And then for any of them where their year 11 mock had felt like a bit of an anomaly, so if they'd not done as well, I looked back to how they'd done in year 10 to see whether there was a disparity. **head of department, selective**

For a significant number though, the greatest concern was around the progress or trajectory of the students, with a need to make adjustments for attendance at revisions classes, the amount of effort made since the mocks, and the general trajectory of the student's work. This is further considered in section 3.2.4 that follows.

Mocks were sat at various times before the lockdown, and so different degrees of correction might have been needed to take account of the additional learning time. Some had been sat far enough back that they would have been less useful as evidence.

> We had done some mock exams back in November for GCSE, a bit less helpful because it was November, but we'd done it at the end of February for A levels. **head of department, independent**

Some centres simply hadn't sat any mock exams, and so had to rely on a more varied set of evidence, including more reliance on class tests and teacher judgement.

> We hadn't done a formal mock with the GCSE this year, because so many of them resat in November. […] So it was just based on teacher assessments, what we'd done in lesson and our feeling as to who was better or worse than who. **head of department, sixth form college**

> Within the school I'm working at […] there are large departments, particularly say our English department, which […] feel overwhelmed by frequent formal assessments. So they don't set many. […] They don't have proper full length mock exams and things like that. **teacher/tutor, academy**

3.2.3 Other sources of evidence

In addition to mocks, a wide variety of class work and class tests were also considered, many of which have been mentioned in previous quotes. These had variable weight when measured against mocks, decided by individual centres, departments or staff. There was a strong tendency, related to the idea of tracking the students' progress discussed in section 3.2.4 below, to weigh the most recent evidence the most highly.

> We put more weighting on […] the most recent marks that we had, we felt that was fair, because it showed that they were getting better and that they were putting more effort in then […]. So we hope that that balanced it out. […] Those who are getting better got more marks, more weighting for that as well. **head of department, comprehensive**

Differences between subjects were likely to have arisen, and teachers and tutors were clear in what were the most important sources of evidence within their subject areas to make judgements. We did not have a sufficient sample to conclude any subject-level differences, and neither do we have to space in this report to reflect all the varied considerations specific to particular subjects. As such, this section highlights the most cross-cutting considerations.

For some GQ subjects, participants commented that coursework (non-examined assessment) was a useful source of evidence. Because of this some centres had gone to efforts to ensure that coursework was in a state to assess. For example, a music teacher commented that they had gone to great lengths to help the students complete their coursework.

> We created a day for them to do that to get all of their work done and that happened to be two weeks before lockdown. So they'd actually completed all of their coursework. So there's three areas, there's an examination and then there's a composition and a performance, that's the three. And the composition and performance, the coursework stuff, was done, in a box, ready to be sent. **teacher/tutor, independent**

A drama teacher described how they had to rapidly put together some evidence that they would be able to use for the CAGs, while recognising that this was unlikely to reflect their best work.

> So the poor kids on the day, when we found out that we were going into lockdown, I said to all of them […] 'right, you need to come in and you need to video these performances now'. And they were like 'but we're not ready'. I said 'I don't care if you're not ready, I want you to video what you've got'. **head of department, academy**

Moreover, this head of English reflected on how their A level literature coursework helped them distinguish between students' rank orders.

> Just purely by luck we had got the coursework in before lockdown. So we did mark it and use it. Not obviously for 20% of your grade like it would be for coursework, […] probably for your fine tuning: two students are a middle C grade, but actually that student did far better coursework, so we might put them a bit higher. […] It was probably for fine tuning rank order that we used coursework more than anything. **head of department, sixth form college**

On the other hand, some participants mentioned that coursework could be difficult to evaluate. This was especially the case when comparing students whose coursework was at different stages of completion. They expressed that it felt unfair to rank a student with partially complete coursework above a student who had fully completed it.

> Is it fair to give the two children the same grade when one hasn't actually done the work, but I'm not allowed to punish them for the fact that they haven't done it? […] I think that was quite difficult. **head of department, independent**

This Art and Design teacher reflected how students would have improved their coursework before submitting it.

> The reality is that you mark it all as if it's finished at the end of February end of January in our case, but actually you say to them you know what, this is handed in in May. […] carry on adding to it. […] Actually they've matured so much in doing that exam process. So sometimes the quality of what they can add at the end can actually really help the grade. So that's the sort of process, so yes nothing was really finished. **teacher/tutor, further education establishment**

However, art and design students did generally have a lot of evidence to consider putting their teachers in a good position with evidence to support their judgements.

> I can pick their best project and elements of another project or two other projects, or even three other projects in some cases. And so I've got loads of evidence of what they're capable of. And by the time we went off school in March they'd also done the drawings and the research for their component two and so I had lots of evidence for their exam project as well. **head of department, comprehensive**

For a wide variety of VTQs, coursework was heavily relied upon and provided a well trusted source of evidence as it was generally completed (and often marked and internally moderated) or at least substantially complete. Therefore, the use of coursework was less of a prediction than a data entry exercise for many. VTQ issues are considered in more detail in section 3.4. However, it is worth noting that the Extended Project Qualification (EPQ) fell into the same category, being project-based and largely completed at the time of the lockdown.

> For EPQ that was actually relatively straightforward, because we'd actually got completed student essays, and also the production log, I could have a good idea pretty much where that student was. […] I have actually been an examiner for EPQ as well though, or moderator, so straightaway I'd got a good idea what sort of grade that work should be at. **head of department, sixth form college**

Participants also commented on the use of predicted grades as a source of evidence. While a great many centres produce predicted grades for all their students every year, and many weighed those grades into their judgements(often through the data spreadsheet approach), sometimes the predicted grades were not considered to be at all useful.

> So in all subjects except English, and just to say this was just for A-level and GCSE, not for vocational, the grades, the teacher predicted grades were a little generous, some subjects more so than others. **head of centre, university technical college**

> We had access to our predicted grades, but we didn't really use them, because we felt our assessments were probably a better indicator of that anyway really. **head of department, sixth form college**

3.2.4 Qualitative considerations

Almost all centres in our sample considered both student progression (their trajectory, whether stable, improving or, rarely, declining) and learner attributes and factored those into their judgements. While different degrees of weight were given to this in different centres, generally, students who showed evidence of more recent improvement had this recognised in some way. Attributes such as their attitude, how they took on board feedback (including their response to their mock exams), attendance at classes (such as revision classes) and their ability to perform in exams were all usually factored in to some degree. Section 4.1.4 contains more detail about specific qualitative factors and how they were considered. This section looks at how and when those factors were integrated into the overall CAG judgements, and used to adjust data-led approaches.

It was very frequently reported that the outcomes from the data process only needed small tweaks based on these other qualitative considerations. There were a wide variety of diverse descriptions of how qualitative attributes were used to arrive at or adjust initial grades and rank orders, which we can only offer a flavour of here.

> You've got students who have maybe got poor attendance or have been dealing with issues or have been slow to engage with the work but perhaps the teacher - I mean that's the bit where it comes down to judgement and knowing the student. If you know that that student's been engaging with additional work after school, if they've been starting to work independently, attending workshops and things, then that might result in the teacher that teaches them saying look I know this guy and they're going to get, they're going to do much better than their mock suggests. **head of centre, university technical college**

> Do we think that student has been working as an average hardworking student? […] So then we used attendance at lunchtime sessions, study sessions, everything that they might have attended. And if they had good attendance and all that sort of thing, and they were obviously making an effort then we feel like: yeah OK that progression is going to be there. **teacher/tutor, sixth form college**

> We know the ones that are going to make an effort and the ones that aren't going to make as much of an effort and using all that information to try and get an overall picture. **head of department, comprehensive**

Not every single centre did use qualitative considerations, some ruled out any subjective opinion and only allowed a data-driven process. There were mixed views

about this. Some interviewees were conflicted about being part of this data-led approach. For example:

> It's harsh to say but you're not treating them as humans in that sense. That's the whole point of being objective, do you know what I mean, you can't think 'oh I know such-and-such really well and, you know, oh she really deserves a 4'. You've got to think 'well, objectively based on the data'. **deputy head of department, comprehensive**

In one instance a teacher felt that the school had actually given poor guidance by allowing attitude and behaviour to be considered as part of the CAG determination process.

> Attitude, it's sort of like the nail in the coffin for some children as well, because for me it shouldn't be about attitude. That's not good evidence, or from what I read [in the published guidance] it didn't mention attitude or behaviour and that's not good evidence. **SENCo, comprehensive**

Some commented that the idea of ruling out teacher opinions was challenging, but added that the data-driven approach could be as accurate as an approach including qualitative evaluations. For example:

> So they asked all subject leaders across the school to say what evidence we wanted to use to build our grades. So my line manager […] put that together and sent it to me and I just was really shocked because it didn't seem to have any input from me and what I know about the kids. […] 70% of it was going to be based on the mock exam grade, which included speaking mocks that we'd done, and then […] a certain percent of it was going to be done on written work that they'd done in class. […independently] I ranked the students based on my opinion, literally on my own gut feeling. They then used this percentage thing that they'd got from the mock and from the unit tests and ranked the students and out of 26 in the class I only got one wrong […] which I was absolutely astounded by. So, just me based on [knowing] the students really well, but that data and that stuff did prove to be right. **teacher/tutor, comprehensive**

There were a couple of cases where the rank orders produced by a data-driven approach did not seem appropriate and the staff had been able to adjust the rank order using their knowledge of the students.

> There was a discussion because we thought well 'hang on, that's put so and so above so and so and that's not quite right'. And then we had the flexibility to say well 'let's go back into those marks and

> maybe that is a bit generous for reading' and 'that's possibly where they're inflated' and then we could adjust that. **head of department, independent**

Sometimes, when it was difficult to ascertain the degree to which a student's performance was going to improve, interviewees used the degree of improvement of students in previous years who shared similar characteristics as a template.

> If people were really struggling with a particular pupil, I said to them 'think of a student last year who got that grade and compare them', that sort of thing, or 'think of a student who was very similar last year and have a look at what grade they got'. **head of department, comprehensive**

> I know that that child will pull something out of the bag on the day, because he is exactly like that child from last year, the child the year before, the child the year before that, etc. etc. They will do it because they're that sort of child. **teacher/tutor, independent**

One message that came through strongly in several interviews was that if any adjustment needed to be made to the CAGs and rank orders calculated in a data-driven approach, there had to be evidence to justify it.

> I suggested a ranking in each case to the teachers based on the information I had and invited them to modify the ranking based on their knowledge of the students. So I felt it was quite important that I gave them the initial ranking based on what the data was showing so that they understood clearly that if they were making a change it needed to be considered and justified, because the information we had suggested otherwise. **head of centre, university technical college**

> We couldn't have anecdotal evidence. Well yes this student we know would turn up to the revision sessions so they would be better. We could only base it on evidence that we had in front of us. That was quite important. **head of centre, comprehensive**

VTQ judgements were much more heavily based on the available evidence from classwork and coursework. Centres offering both GQ and VTQ commented that there was usually a little more judgement or prediction required for determining CAGs and rank orders in GQ.

> I think it was more flexible with GCSE because we were saying 'had this learner continued at this level, what would have been their outcome?' and I felt like there was room there for positive progression. But in terms of FS [functional skills] is very much like if they didn't do it then it's probably going to be a fail. […] I felt like our judgement was valued a lot more for GCSE, whereas FS was purely evidence [based] and, like you say, data driven. **teacher/tutor, further education establishment**

Nevertheless, in putting together centre judgements for VTQs, some interviewees commented that they also made use of their knowledge of the students to make more qualitative judgements and adjustments to their submissions.

> Because we were able to talk about them as a person, how much we knew them and therefore why we were making the decisions, we actually got commended [by the quality department] for the fact that we knew the learners so well. So it was holistic, but integrity wise we wanted to have solid evidence to support why we were making the decisions we were. **head of department, further education establishment**

3.2.5 Different degrees of teacher input

There was a range of decision-making individual class teachers/tutors had. It was most frequently reported that the class teachers/tutors took primary responsibility for making judgements using the kinds of evidence discussed above. However, there were a number of cases of class teachers/tutors being presented with a data spreadsheet which included a provisional set of grades and/or a rank order and being asked to review or simply approve this. More frequently data was passed to the teacher/tutor and they were asked to use this as their starting point, but were expected to use their knowledge of the students, and any additional evidence, to refine this list.

In several centres, while the teacher/tutors initially determined their own list of grades and rank orders, they were then presented with a data spreadsheet and asked to review their initial judgements in light of the data.

> We had a first draft submitted to myself and I modelled that data. In all cases that was sent back to departments, because it was inconsistent with previous years and departments were asked to review their choices and select a critical friend within their own department and to critique each other's grades. They weren't told to bring them up or down, they were just told to do a stress test on those grades. **deputy head of centre, further education establishment**

In a small number of cases we saw a parallel process where a data spreadsheet was compiled at centre level or by the head of department (complete with full ranking and grade profiles), while at the same time the teacher/tutors made their own judgements. These two sets of outputs were kept separate until the final agreement process within departments.

> We hid the actual data-official grades. We then asked staff to make their own judgements, so they weren't influenced by that data-driven grade, because we didn't want them just to [make a] decision off assessments that kids potentially hadn't worked for, they had different opportunities for, whatever it might have been. **head of department, academy**

There were also different degrees of class teacher/tutor involvement in the iterative process of refining the grades. Mostly, our interview sample were involved throughout the process of agreeing grades at department level as described in section 3.3. However, it was less frequently reported that teachers produced an initial set of CAGs and/or a rank order and then passed these to more senior colleagues with no further input.

> I think they sort of felt that if the teacher was involved in the moderation of the grades that would potentially result in some bias. So they, I don't know what happened after that, they basically said thanks for submitting these, these are now going to be checked by the line manager and we're not going to tell you what they were. And so I still don't know. The only thing that I heard from my line manager was that I know that they were looking at the previous three years of results and trying to keep the results in line with that. **teacher/tutor, sixth form college**

> And the exams officer just basically asked for the grades and that's it. There was never really a conversation with them. **teacher/tutor, academy**

3.2.6 *Specific considerations around rank ordering*

In our interview sample, teaching staff indicated that a data-driven approach often gave a rank order that was used as a starting point. Following this, grades were allocated by staff judgement, or by the use of an expected grade profile (often based on prior centre performance), with some adjustment around grade boundaries. For more teacher judgement-led approaches, it was more common for grades to be assigned first, followed by ranking within the grades.

Rank ordering did generate some stress for teachers and tutors. This was partly through the basic difficulty of ranking diverse types of students and partly through an awareness of the potential impact their rank could have following statistical standardisation. From the interviews, it appeared there was more time spent on the rank at grade boundaries and less worry about the rank in the middle of a grade. This was probably driven by the expectation that statistical standardisation would be more likely to, for example, adjust the grade for students at the top or bottom of the rank within a grade.

> When we got towards the middle of the grade we were a bit more like, it wasn't like in fine detail. It was like well, you know, they look quite similar let's just put that person ahead of that person. So it wasn't quite as forensic as the grade boundaries where we were conscious that there might be people moved up and down. **deputy head of department, comprehensive**

> It was mainly the borderline ones, the ones between each grade, where you had some who were 8s and some who were 7s, and we'd discuss all of those pupils around that overlap grade and then we'd move down to the next one. **head of department, comprehensive**

Another significant difficulty was rank ordering students across different GCSE entry tiers, where both lower and higher tiers could access the same middle grades, but they had been taught separately and taken different mock papers. A variety of approaches were taken to arrive at the fairest ranking possible – one department used the common items across the papers to help.

> Because there is a crossover on foundation and higher tier, the questions, so we could look at the [answers to the] actual questions. […] And so if they'd taken a higher paper here's where they would be and if they'd done the foundation they would have got all these marks, so a mix of that really and just knowing the students. **teacher/tutor, comprehensive**

As we have seen above, rank ordering could be completed based mainly on numbers in a data spreadsheet, with varying degrees of teacher/tutor judgement and discussion factored in. In cases where a large cohort needed to be ranked, data was the main way of ranking similar students who had already been given the same grade, particularly where the students were taught by different staff.

> I ranked the whole year group based on what people were saying [for their grades] which then of course would create clusters of students say maybe from position 30 to position 60 might all be on exactly the

> same grade, so they needed sorting out. And I tried to sort that out using the mock data that I had, because that was the quantitative data. **head of centre, university technical college**

> How do you know that a student that you're saying is a grade 4 is better or worse than somebody in another classroom that is saying they're a grade 4. So that was where the objective pure data evidence came into it, mock scores. **teacher/tutor, comprehensive**

Some centres thought that having to rank order students was a good exercise as it forced teachers to justify their grades more.

> The rank orders are probably where we had an awful lot of our discussions, because not only did it force you to justify again the grades you'd given them, but if meant that teachers were really thinking about who the top end kids of that grade were, who they thought actually they were quite borderline. **head of department, academy**

3.2.7 Challenges faced

Most of the challenges the participants discussed related to many of the issues covered in the qualitative considerations section. In general, students with uneven levels of engagement and effort were difficult to place in grades and rank orders. Taking into account the longer-term performance of students who had under-performed in their mocks, sometimes through particular difficult circumstances or just through bad luck or lack of effort, were frequent concerns.

Another problem mentioned was how to make fair judgements when students could invariably underperform in exams, and in unpredictable ways. For instance, reductions in performance due to stress or anxiety, or only partially completing the exam. There were a wide variety of specific cases described in detail. In fact, most interviewees mentioned at least one difficulty in their interview, and it appeared to be rare for everything to be straightforward.

In every case though, these kinds of difficulties were the focus of discussion between teaching staff, and participants reported doing their very best to ensure fairness for every student. Sometimes there was no other way than to rely on the data. In some other circumstance, an adjustment was made, but, with the awareness that the decision had to be fair to the rest of the cohort. This was driven by the understanding that placing a hard-to-place student near the top of their potential performance range could potentially see a more solid student receive a lower grade than their CAG following statistical standardisation. In this section we describe a few specific

categories of learners and some logistical issues, rather than these individual student cases.

Determining CAGs and rank orders for private candidates was seen as problematic for most centres that discussed them, due to the lack of familiarity with their work. Sometimes, a lot of the private candidate's work had to be reviewed, and sometimes re-marked, in order to determine the rank and grade.

> It was quite difficult because obviously, the student had got a private tutor who has clearly said to him a particular grade and has plastered that grade over all of his practice exams that were submitted to me as evidence, and I didn't agree with that grade. […] I looked at it and then I asked another member of my department who also teaches the same, who teaches English language A level, to go and look at it. So it was good that we could then have that discussion between the two of us. **head of department, independent**

One centre reported setting private candidates tests that had been sat by their own students in order to help rank them.

> We made him take some tests as well that we had made our own pupils take so that we could calibrate how he got on compared to them. **deputy head of centre, independent**

In some cases, centres had decided they were not able to make judgements for private candidates who they did not know. They reported fearing potential detriment to their own students by including them in their rank order following statistical standardisation.

GCSE English language and mathematics resits were a common issue for centres, particularly further education establishments, with many students to rank who were all performing at a similar level, usually close to the vital grade 4 boundary. This was often a binary choice for each student between grades 3 and 4.

> Primarily we take the students who got a 3 at GCSE, so we were predicting grades […] for a bunch of students who almost all of them got within a few marks of each other last time they sat it in summer. So we're trying to really split hairs between students who are very similar. That was quite complicated. […] The middle 60-70% of the cohort, they're somewhere between a middle 4 and a high 3, and trying to choose between them was very hard, particularly when, like I say I'm trying to work out this student that I taught, are they better or worse than that student who I've never met, who that teacher's telling me was at a similar level. **head of department, sixth form college**

> The centre assessment and ranking within English and maths was a challenge for us, because they're a large cohort where we have around 1,000 learners doing GCSE English and almost 1,000 doing GCSE maths as well. […] And so ranking was a real challenge when you're looking at large cohorts and multiple groups on the same qualification, all the different teachers and the concept of […] ranking for the teachers, that they can rank their students, but how do they rank with the other 20 groups of students with other teachers? So there were real intellectual, pragmatic, operational challenges. **senior leadership team member, further education establishment**

One area where a few centres sometimes expressed a lack of confidence was in awarding top grades, particularly where they did not historically have that many students achieving those grades.

> [One difficulty was] feeling confident to be able to award someone a 9 and an 8, they're probably, you know, what's the difference between an 8 and a 9? You know, sometimes that's about performance on the day. […] So an 8 and a 9 is very difficult. And I think that we maybe underestimated some of our 9s. We don't get lots of 9s but we get some and maybe more of them would have got a 9. **head of centre, comprehensive**

Although largely it seemed that centres had everything to hand to make their judgements, sometimes pieces of evidence required to make judgements were not available due to the lockdown. This could be due to difficulties connecting to centre IT systems from home, physical work being stored in the centre, and there were even cases where centres had given back work and mock exam papers to their students because they thought exams were going ahead despite lockdown.

> As soon as we began to think school would shut here, but not cancel the exams, as soon as that began to be on our radar, we gave all our children their books and their mock papers and everything. So we didn't actually have any exercise books or paperwork in school to really look at. **SENCo, comprehensive**

The practicalities of getting back into closed centres was mentioned.

> It took me two weeks to get permission to go back into the college to get material out that I needed. […] I had to get back onto my desktop in my room because that's where the spreadsheets were held. You know, it was just ridiculous. I had to get back into classrooms and come home with rafts of books. **teacher/tutor, further education establishment**

The general limitation of working remotely without student work in front of everyone was repeated.

> It is difficult when you're on a Zoom call and you're at home, because you don't have access to everything that you need at home, because everything is in school. And it's not as easy to sit down and talk about different students and look at where different students are when you're at home and you've not got access to the student's work. **head of department, university technical college**

The difficulty of using remote meeting software to have discussions was echoed in quite a few interviews, but no-one reported that they felt it had any substantive effect on their judgements. Largely participants reported just getting on with judgements despite the difficulties and stresses they faced.

> Although it was strange, obviously not being in the same room as each other when it was finalising things, I think we were all quite comfortable with each other doing this. **teacher/tutor, sixth form college**

> Just working on a 12 inch screen is really tough. I had to get my TV on the kitchen table one time, so I could have two screens […] and I could just look at all this data. […] Another thing as well I've not had my dining room table for months, I've just about got it back now during the summer holidays, but I've just been working completely on there, and my kids have been working on there as well though. So they've demanded the computer as well, we've only got one laptop in the house. **head of department, sixth form college**

Where there was shared teaching of a class, and a need for the two teacher/tutors to agree their judgements, again no-one reported major difficulties agreeing. Most were able to agree everything after discussing any differences in their views, while less frequently a more data-driven compromise of the grade or rank order was agreed. Where tension or dissatisfaction was reported this tended to be more at the departmental agreement stage, which is described in section 3.3.1.

One final difficulty to note was that of the absence of teaching staff due to them being unwell (given the pandemic situation in the country), or having left the centre (an unavoidable aspect of normal staff turnover). Again, this was just something centres had to cope with as best they could, either trying to contact the person who had left, or simply managing with what they had.

> And the people who do leave, and they've been here for maybe two or three years, have an attachment to it and will always go above and

> beyond and show that professionalism, even though they're not employed by us, will actually support us and help us. That's the culture that you've got to have. **head of centre, comprehensive**

This could obviously have impacted on the accuracy of the judgements in some cases. One mitigation would have been to rely on data, including checking judgements for students across subjects. For VTQ judgements the most widely reported challenge was lack of time, sometimes due to relative lateness in decisions or guidance from AOs. This is detailed in section 3.4.

3.2.8 Giving the benefit of the doubt

The pattern of grades and any optimism or grade inflation in the final submitted grades is discussed following agreement within the department or centre in section 3.3.3. In this section we will only consider the issue at the initial class-level judgements.

A very common opinion was that the students were much more likely to under-perform on an exam – 'have a shocker' or 'crash and burn' as they said, than over-perform. Interviewees commented that it was impossible to know which of the students who had the ability to get a grade would have under-performed on the day. Therefore, teachers tended to give them the grade they had the potential to achieve on a good day.

> People have a shocker on the day, but you don't know who they're going to be. So I think it's difficult when you've got a group of students and you think these four students, they're not all four going to get an A, but all four of them could get the A. […] so it seems a bit unfair to give the fourth one down a B. **head of department, independent**

> When kids get their grades you always get that odd one who proper crashes and burns. And of course you don't predict that.
> **teacher/tutor, comprehensive**

> For those students who are genuinely borderline between two grades, [they] could go either way; whereas in an exam roughly half of them will get the higher grade, half will get the lower. That's where it's really hard as a teacher, because how would you pick which half you're saying right no actually I'll give you the lower grade. So I think that's where if I'm honest we've probably slightly over-predicted. That's where I'd imagine a lot of places have over-predicted. **head of department, sixth form college**

One respondent summed this view up well with a golfing analogy.

> I compared it to playing the perfect round of golf, which I don't play but you understand how it works, in that when they're trying to determine your handicap they look at the ten rounds you've played and say look what's the best score you got on the fourth hole, what's the best score you've ever got on the eighth hole and if we add those together you ought to be able to do a round of X, but you never do because you never get every hole right every time. **deputy head of centre, independent**

Except in cases where the class teaching staff were fitting their CAGs to a pre-determined grade profile based on prior years, there was a widespread recognition that their CAGs would be slightly higher compared to previous years, but that this was perceived as being absolutely fair and the right thing to do for the students. Teachers and tutors felt certain that it was entirely inappropriate to guess which students would underperform when there was no evidence or reason to make that judgement.

> I had a very borderline class and I would not […] give anybody a 3. All of them had a chance of getting a 4, all of them. Some stronger than others, but I would not put anybody at a grade 3, because morally I couldn't do it. […] If there was a child and an English or a maths teacher was saying they had a chance of a 4, I think we are duty bound to give them a 4. That's a really strong opinion of mine, because the effect on their lives and the nonsense of doing a resit in November after not learning since March is grossly unfair to them. **SENCo, comprehensive**

> I think a slight over-prediction, but I get where that has come from. I think that will be teachers who in good conscience have gone for totally fair grades, who haven't of course legislated that one student in 10 who bombs on the exam, or gets a grade lower. So I think that will explain that. **head of department, sixth form college**

3.3 Final submission agreement and quality assurance process

This section looks at how centres managed the agreement process for the final submission of CAGs and rank orders to the awarding organisation(s). In the analysis that follows we largely separate the agreement process that occurred into two stages. The first includes agreement within departments, often whereby the initial qualification judgements were made and sent to the senior leadership team (SLT) for

sign-off. The second is the agreement process within SLT at centres. It is important to recognise that these two levels of discussion usually overlapped, with either direct discussion between departments and SLT, or the judgements being passed back and forth during the review process.

First it is worth noting that in the majority of our interviews there was a great deal of discussion around the final agreement of the centre judgements, but, this was not always the case. As we saw in section 3.2.5, some teachers or tutors worked alone to make their judgements, and then passed these direct to a line manager or SLT who finalised the submissions without any further input from the class teacher/tutor.

While this independent judgement process might appear to be more likely to occur in subjects with only a single teacher, in fact in our interviews, discussion and checking appeared to occur with line managers and SLT in these cases as frequently as in larger departments. The above approach appeared to be more centre- than subject-specific.

There were also cases where there was less discussion at departmental level. This particularly occurred in centres adopting a centralised data-driven approach. Often, there would just be a check by departmental staff that the grades sent from SLT did not look unreasonable, sometimes with little discussion of individual students between staff in that case.

> I wasn't in charge of submitting all the grades; I was just consulted about my class. So from my perspective I was given a set of grades and said do you agree? […] Basically I was shown a spreadsheet and me, […] the head of department and the teacher of the other class, we tweaked a few things. But they didn't change that much from what we thought. Nobody changed more than a grade sort of thing. So and then it was sent off and I didn't see it again. **teacher/tutor, comprehensive**

However, discussion between department team members was the most commonly reported approach.

3.3.1 Agreement at a departmental level

This section describes how departments consisting of more than one teacher/tutor worked together to agree and quality assure grades. A key consideration in all of the departmental-level agreement processes was the need to merge students from different classes into one set of judgements, and this was particularly challenging around the rank order. A strong contributor to merging rankings across classes was the use of common mocks or standardised assessment tasks across classes. These kinds of common assessments, as well as the use of comprehensive data files,

allowed heads of departments, or departments as a team, to agree a common rank order without relying on opinion too strongly.

> The statistical data that we used, everybody had done the same thing, so it wasn't like my class had done one test and the others had done another test, but we were using the same, for the same mark. They'd all done the same. We only used data where everybody had done the same thing [to rank order]. **head of department, comprehensive**

In many departments class teacher/tutors provided grades for the classes they taught, and sometimes but not always, a rank order, which then needed to be combined through some departmental-level process. Sometimes there was some pre-checking by the head of department, looking for anomalies.

> But I did have that consultation process whereby I sent them in and then [my head of department] said have a look at these, have a look at these marks and have a think is that really a fail, would that be a pass? And yeah, we did do that process. **teacher/tutor, further education establishment**

> And sometimes you found high ranks of students with much lower performance, and you identify that that was teacher bias, and you went back to the teachers with a new suggested ranking, and bar one case and one student they all agreed on it. **head of department, further education establishment**

Frequently the head of department checked class level grades against previous years' data.

> Last year our value added in science was plus 0.28; my gut feeling was that we were heading for a value added of about plus 0.5 this year. So whilst I felt we'd gone through a rigorous ranking process - so I felt the order of the kids was sound, when I put the first round of predicted grade data through [the] SISRA [analytics service], it came out with a value added of something like plus 1.6. Which I just, you know, I knew was obviously too high. **head of department, academy**

Sometimes (only occasionally) the head of department made the initial judgements, which were then discussed/checked by teachers

> So the head of department did the GCSE ones and then we all got involved in the discussions checking that we were happy with [the grades]. So then we did that for our own classes then checking the

> rankings and comparing. So some of that was individual, some of that was collaborative. **teacher/tutor, selective**

Many departments worked as a full team to agree the CAGs and rank orders, either through a lot of smaller discussions or one (or more) long full team meetings where individual students were discussed. These kind of discussions covered the range from decisions being highly influenced by data, to discussions based more on judgement.

> There were three of us involved overall. […] There were some cases where we didn't agree about the rank order, about the grades to award. So we were led much more by the data on that one. So we looked at, we got our rank order eventually to a point that we were happy with before we put any grades on. […] The top three would get As, the next five would get Bs or whatever it was […] So we basically went through and went right, to be in line with the previous years, […] this mapping of grades that would give us a value added of zero. […] We just found it was easier, because there were that many disagreements between [the teachers]. […] It just felt like the quickest and most rational way of doing it really.[…] I think I led the data driven one, because I felt that the other teachers were perhaps over-predicting a little bit when we talked about the gut feeling grades. **head of department, sixth form**

> I've got two other members of staff within my department, so their classes were then shared with me and we had a discussion around 'why have you given that student that grade? I don't agree with that grade. Can you tell me why you've given that evidence?', that sort of thing. And then […], we made any adjustments that we needed to make and then we shared it with the […] head of centre. **head of department, university technical college**

However, some did not produce the department's set of judgements through group meetings. In several cases the CAGs (and sometimes the class rank order) were done by teachers on their own and then the head of department combined them and created the rank order alone.

> So for the A level chemistry, I've got two groups, two sets of 24. I assigned two rank orders of 24. […] And that's probably where I felt sorry for my head of department, because he had to basically take [all of the] classes, and put [the students] into a rank order of 440, that was really tough, and that was a nightmare for him. **head of department, sixth form college**

> We asked those teachers to grade and rank them individually without consultation with one another, and then I took that information in together and collated it and they were almost spot on. […] They ranked within class, then I did the ranking within the grade. **head of department, comprehensive**

In the case of a small centre, the head of centre might share this role with the head of department.

> It started off class by class, where class teachers would rank and then they would do the predicted grades. But then obviously, we had to combine those. And at that stage where they were combined it was myself. So taking science as an example again, I worked with the head of science to have a look at those rankings, so that we were looking at them across the whole year group. **head of centre, university technical college**

In one interview it was clear that the lack of collaborative working at this stage had undermined the teacher's confidence in their judgements.

> I would have liked to have done it collectively. I would have liked to have met as a department to do it, especially because there was another biology teacher who had separate classes to me. It would have been nice to look at how she was doing it, whether we were doing the same things. […] I had no idea whether the way that I was doing it was similar or different to other people. **teacher/tutor, comprehensive**

Sometimes the class teacher/tutors were not happy about their limited involvement in agreeing the final departmental judgements.

> Once you'd given your grade and fine-tuned it, you rank ordered your class, and then that went to the subject leader. And the subject leader then, so for my subject had eight different teachers and all that data and somehow or another she had to decide what order it was. That will not be accurate without a shadow of a doubt, because she couldn't speak to all of us at the same time in a room, or in a meeting. […] None of us had all the knowledge of all the children. **SENCo, comprehensive**

> I was not involved in the ranking, so I wasn't able to put my case forward if someone said 'we're moving this child out from say a 6 to a 4' for whatever reason, because we need to fit in with our figures, I didn't have my chance to say 'no, in physics I feel they would have

got this and with the wind in the right direction on the day they may have got that'. **teacher/tutor, comprehensive**

One teacher was particularly upset and felt they had been ignored.

[The head of department] changed them and she sent it back and said well it looks like this, are we all happy with it? And I basically said no not really. And she said oh well that's what I'm going to submit. […] If the person is not amenable to listening to that, then there's nothing you can do. **teacher/tutor, independent**

However, more usually the agreement process of the final set CAGs and rank orders was run by the department and agreed between all members involved with producing them.

And then all five classes came to me as head of department, and it was my job, somehow, to merge the five ranks into one whole rank before I then put that out to the department. And we had one meeting with all five of us on Google Meet. I shared my screen and we had that spreadsheet up and we worked our way down it. **head of department, independent**

[The teachers] were free to populate the spreadsheet themselves with their predicted grades and their ranking within the class. But then I overarched that and standardised and developed through conversation with them a ranking that sorted things out between the classes. […] The final thing was for me to once again sit down with the team and say 'look I've done all these things this is now what it looks like'. So we looked again at the rank, we looked again at where the grade boundaries were falling within the ranking and actually, […] there was no conflict within the team; we felt we'd actually done a fairly thorough process. **head of department, academy**

At department level it was clear some hard decisions had to be made. Although generally the process sounded to have been reasonably harmonious, or at least professional, sometimes they could become difficult, reflecting the staff's awareness of what was at stake for the students.

Yeah, there were a few heated words at times. It was really difficult because obviously […] from a class teacher's point of view you want to do your very best by that pupil, you didn't want to give him any disadvantage, particularly for those that you know probably haven't done as well in the mocks. […] And at the end of the day the final say

wasn't left with me, I could only argue so far, but I didn't actually have the final say, which was difficult. **teacher/tutor, comprehensive**

3.3.2 Quality assurance and moderation processes – senior leadership involvement

This section describes the parts of the agreement process where the senior leadership team (SLT) were involved.

3.3.2.1 SLT querying judgements with departments

In many of the interviews it was clear that there were various quality assurance processes put in place, and in some centres, particularly the larger ones, they were often very detailed.

> Obviously, we then had discussions with department team leaders or the middle managers and then they gave their opinion on it. And then we had discussions with the heads of departments and then the discussions with their bosses, then the assistant principal and the head of quality as well. So everybody kept checking it. And then after that point it went to standardisation within the department and then it went to standardisation at college level as well. So it was checks and checks and checks yeah. **teacher/tutor, further education establishment**

The example above is perhaps at the extreme end of the scale, but it is clear that a wide variety of checks and balances were put in place by centres. Sometimes SLT checked the appropriateness of the evidence used by teaching staff to make their judgements.

> So I just then QA-d the evidence basis that we had to [support] the grades that they'd put and I'd put in, while they were doing their individual class grades. […] We then came out with, with sort of an interim grade if you like. And then once we had that grade we went back to teachers to say 'these are what grades we are suggesting the students are awarded' and gave them the opportunity to sort of come up with evidence or an argument to say actually, why they were or weren't [at that grade]. **senior leadership team member, academy**

This evidence checking did not appear to be universal though, in some centres the check focused much more on the "reasonableness" of the grades - from the interviews it appears that almost always the judgements from departments were centrally checked against previous years' data. This was achieved using internal

data, a variety of database analysis tools/reports and external resources and organisations. This check was consistent with the information released by Ofqual about how centre judgements would be merged with statistical predictions to standardise outcomes. Often, the result of this analysis was returned to the departments to be reviewed. There were varying degrees of firmness in this recommendation, from a directive to strictly match prior performance, to more flexibility to accept some increase in grades. It was then up to the departments to re-evaluate the students or defend their initial decisions.

> SLT then communicated back to the subject leader and there were a number of children that we were asked to look at again. […] I was sent a list of three of my class with what I'd put and was asked to confirm if I was happy with that. So at a high level they must have looked at some and then queried it and then asked the classroom teachers to go back and have another look, I think is what's happened. **SENCo, comprehensive**

> And this is what SLT was saying to us. […] If you've been giving a student a 4 for two years, and then you suddenly give them a 7, questions are going to be asked. And SLT were really good about that and they weren't pushing anybody to change, but they were saying obviously questions are going to be asked if you've given them a 4 all across, and then you suddenly give them a 7 or a 6 somebody's going to ask. **head of department, academy**

> I know that [SLT] were working back from the pass rates from previous years. So if it was massively over that we had to look again at it and try and get it in line with the general outcome of the college, which again I thought was, it's fair enough having to do that. **teacher/tutor, further education establishment**

Several head of departments were happy that SLT has been relatively flexible.

> When [SLT had] got everything in, if there were any anomalies or anything they questioned, they got back in touch with people, but on the whole they allowed us to make that decision and then they looked at it afterwards. If we felt we could justify [that] someone deserved a grade, then they were happy to support us with that. **head of department, selective**

Sometimes this type of interrogation by SLT was viewed positively by the departmental teachers, as a way to increase objectivity and reduce bias.

> So we ranked them initially, graded them and then there were some very tough conversations about 'well are they really above that person'. I think because the SLT member and the head don't really know the students, they were able to be that objective voice going 'well they've got this on their mark and why have you put them there', and lots of questions that made us really drill down into our thinking about where they were and why. **teacher/tutor, comprehensive**

However, this checking of grades against previous years' results by SLT was not universally welcomed by class teacher/tutors.

> So what senior management were doing was coming back to staff and saying actually, given your value added statistics from last year or the last three years, you're too high still, knock some grades down. And that caused, I know there was genuine tears among some staff. There were concerns about that. **teacher/tutor, academy**

It was not always the case that departments had to lower their CAGs. Sometimes this check offered an opportunity to increase them.

> The element where when I did the final checks, where there were anomalies and I looked at them in both directions, so I went back on a couple of occasions and said do you know what your 8s and 9s are still quite low, is that right, have you done that because you just think this is right for the cohort or is there anybody that is right on the cusp that you want to give the benefit of the doubt to, because actually you haven't chosen to do so, so far. And again it was 50/50. A couple of cases they said 'look, these guys just aren't quite [as] good as last years' data', and in other cases they said 'do you know what, we're right on the cusp, so yes if I can put two more through onto that grade above we will'. **deputy head of centre, independent.**

It was usually within SLT that comparisons were made for individual students across subjects to flag up anomalies.

> Presenting a teacher with a series of grades across subjects and saying 'are you sure that in English this student's only going to get a grade 3, because in geography and in drama they've been awarded a grade 5, it seems unlikely they'd get that if their English was at grade 3, can you justify why you're giving them a grade 3 for English'. **senior leadership team member, academy**

It is also important to note that there was not usually just a one-off review by SLT, there was usually an ongoing discussion. The following quote is typical of many of the interviews, where there were several rounds of discussion.

> Our process wasn't we did the grades and we send them off; […] There was quite a few comings and goings. So if SLT saw something, they'd come back to the classroom teacher and ask their opinion on why is that. It was very much a consultation process in our school. **teacher/tutor, academy**

3.3.2.2 Types of final agreement processes

Centres needed to have some way to decide the final CAGs and rank orders that would be signed off via the centre declaration form and submitted to awarding organisations. This section describes the kinds of process SLT would use to finalise the submissions.

Sometimes, following the compilation of the initial class-level judgements, the final grades and overall qualification rank order was decided by SLT with no involvement of departments.

> And then once we had the grades secure, we went and ranked the students within the grade. And then to do that I used, again the same basis of evidence, just making sure that I considered the mock scores primarily, followed by the assessment and obviously their teacher grade within the set. **senior leadership team member, academy**

Quite a few centres held one-to-one, or small meetings between the head of department and one or more representative of SLT to check the judgements. Sometimes this included a comparison to previous years' results, sometimes a check on randomly selected students, or sometimes a more standardised set of questions for the department head to answer. This approach seemed to be more common in smaller centres, presumably with more direct contact between departments and senior leaders.

> I sent [the head of centre] the grades, then we had a Zoom call where we talked through, she would just pick out random students and say 'right what's your evidence for this grade for this student, where's your evidence for that student, this student has performed better across the school than you've predicted, why is that?' And any changes that we made on the back of that was a discussion between the two of us, it wasn't that she made the changes and then told me, it was a discussion between the two of us. **head of department, university technical college**

> You have to have a sign-off meeting with every subject leader with another member of your senior leadership team. You have to go through and interrogate them and we gave them a script […]: 'how did you do this, what were your problems, how did you deal with that problem, is there anything you still feel unsure about?' And just getting the subject leader to narrate their story. **senior leadership team member, academy**

Less frequently members of SLT were present in departmental meetings, so that there was a cross-section of seniority, from SLT down to class teacher/tutors.

> The final conversations comprised of any members of the team involved in that cohort. So [for] the combined science students, we had any combined science teacher along with our head of department or second in department, and two members of SMT, so management. And the idea of that was we went through student by student, every single student in the cohort, discussing the grade we've given them, justifying why we've given it to them and, where necessary, staff would challenge. **head of department, academy**

Larger centres such as further education establishments held more formal awarding boards with a final review of the judgements.

> For A levels and GCSE, as head of quality I was able to implement a centre approach, so holding award assessment boards, producing and modelling data on previous years' attainment, paying particular attention to protected characteristics when we model that, being able to send that back to the teams before [the] award assessment board and saying 'no, it's not good enough, this isn't right'. **deputy head of centre, further education establishment**

There was a general emphasis across all interviews that the centre judgements were a shared endeavour, that they should not be determined by one single person and then submitted to AOs.

> One of our key things is that nothing was done in isolation in terms of views of an individual teacher. There was a discussion about your groups. So your head of department who may not have taught that group or not known that group, when they were looking through they had a discussion and they probed and had some challenging questions. So, you know, why is that student ranked higher than this student. **deputy head of centre, comprehensive**

One of the directors did come around midweek and we sat in the garden and went through everything because I was like I really, you know, I don't want to do this just with one person looking at it. **head of department, Training Provider**

3.3.2.3 Adjustments made to original judgements within senior management alone

As we have seen, it was usual for departmental staff to have sight of the final submitted grades, although sometimes this might be limited to only the head of department. However, sometimes the final adjustments, which were usually based on a comparison to previous years' results, were made by senior leaders who did not pass these final grades back to departments.

I don't know the exact grades. I got the impression that, well I asked whether mine had changed and I think mine were pretty much as I'd submitted. I think some of them may have been changed by managers and I'm not sure they were allowed to tell me anyway. **teacher/tutor, further education establishment**

I submitted [the judgements] to the head teacher. And then across the multi-academy trust they did, they were involved in some sort of wholescale, large scale […] modelling. And then they were going to adjust them. But I don't actually know what they came out as, so my bit ended when I gave them to the head and then whatever they came out as the head signed off on. **teacher/tutor, comprehensive**

This head of centre thought that this was a good approach to take, as it took the weight of responsibility off of the teachers.

It would be me that had the overriding decision. And I also felt that that made it easier for them. I didn't make significant changes at that final stage, but there were one or two changes that I did make, but nothing particularly significant. […] I think they were glad to feel that I was taking that responsibility, so they had done everything that they had done. […] It's very difficult as a teacher when you've got that relationship with that student, when you think they might do it, they might actually be able to do it on the day. So you are going to err on the positive side. And I think with the database that we used, and although I know the students very well, I was one step detached from that and I was able to be a little bit more scientific about it. **head of centre, university technical college**

In one or two cases teacher/tutors recognised that this kind of adjustment would have happened and were resigned to this.

> I don't know what we ended up with in the end, because I don't know how our grades were changed after. […] I gave up, my colleague was quite insistent on trying to find out what had been done and I just said there's no point because they're going to change them, the exam board's going to change them, what's the point, we've done our bit. So I presume we ended up with some E grades, which is probably deserved. Hand on heart they probably were going to get E grades. **teacher/tutor, further education establishment**

More frequently class teachers/tutors were not in favour of this approach and felt a lack of control over the process.

> Once the grade was submitted to SLT, I don't know the trail after that, because it wasn't part of my role to know the trail after that, if that makes sense. […] The grades were changed, so later on […] we all had a briefing with head teacher, and [he] said to us that he had changed grades looking at the statistics of the school over the last few years and obviously what Ofqual were requiring him to do. […] He didn't change the rank order, but he changed the grades. I'm not a subject leader, so whether subject leaders know more than I do I don't know, but the grades were then changed. So subject teachers who allegedly know the children the best gave a grade and then were rank ordered by subject leaders who don't know all the children and then they were changed. For me they are completely fake fabricated artificial grades now. **SENCo, comprehensive**

> When I looked at the final sheets and the grades, predicted grades, I don't even know where those numbers came from, because they were not the numbers that I had ever talked about. And there was another student on there who had been completely downgraded and I have no idea why, where that number came from for her, and I'd already picked it up on a previous version and it hadn't been dealt with. […]There was a real process going on, quality were involved, the exams office were involved, you know, so you had to check them individually to be certain that the right thing was happening. And in the final analysis it left us and it went somewhere else and again, I don't know [the final grades that were submitted]. **teacher/tutor, further education establishment**

It wasn't unusual for class teachers to express unhappiness at the basic idea of using performance in previous years to moderate grades.

70

> I know in the back of the mind of the SLT is that they needed to keep their Progress 8 at around zero. And I think that played too much of a part in how ultimately the grades were awarded. I think there was a couple of times they felt that they were too high above the Progress 8, so they went back and started looking. **teacher/tutor, comprehensive**

A final point to note here is that often the processes and meetings in which SLT were involved were frequently the point at which access arrangements (including reasonable adjustments) were applied to individual students. This is described in detail in section 4.1.5.1.

3.3.3 Final submitted CAGs and rank orders

This section describes observations around the final submitted CAGs and rank orders.

3.3.3.1 Matching previous years' results

Carrying out a check at some stage on the judgements compared to previous years' results was almost universal in our interviews. Even if this had not been done early in the process, such as within departments, a check and sometimes an adjustment of the CAGs was carried out as part of the final check within the senior management before the centre declaration form was signed and the centre judgements were submitted, as described earlier.

The adjustment of CAGs stemmed from several causes. There was full awareness of the existence of the statistical standardisation that was to be applied by the exam boards, although not always total clarity about how this would operate. There was evidence of some 'second-guessing' of how standardisation would work and this featured in how the quality assurance worked. There was also evidence that some centres did not feel confident about what adjustments would be like, which made some want to avoid statistical adjustment.

The general pattern though, despite internal moderation, was that centres were entering CAGs that were typically slightly generous. We saw evidence that some centres were hoping there was a little bit of leeway, or 'tolerance', applied to generous CAGs before statistical standardisation would be activated, and they were attempting to find the right balance between optimism and triggering this standardisation.

> We've gone very much with a tolerance level of about 5 to 7%. History will have improved by 5%. If it had improved by 15% in 9 to 4s, you know, that data would have been questioned. **deputy head of centre, comprehensive**

> And we weren't looking for something that matched exactly; we were looking for something that was within a level of tolerance I suppose. […] If we're talking about a level of tolerance, so I'm not saying you can't put them in more optimistically, we know it's a more able cohort, we know you've worked hard on it, but it was just, the gap was too big. **head of centre, comprehensive**

> I think we said, you know, I think the unofficial rule in coursework is you want to be 5% high, OK on the basis that you want to be optimistic, but not so optimistic the exam board will bring you down. […] So I think that's where we want to be with it, because clearly we want the best outcomes for our young people. **head of centre, university technical college**

In a very small number of cases, department-level CAGs had been a little bit below previous years' results (although note that often this comparison was made to just the previous year, and not necessarily to the full 3 years of results that it had been announced the statistical model would take into account for most GQs).

> We then submitted our grades to the data and senior management team. They also told us what our […] level 3 value added would be and mine for this year would be lower than last year's. So he said did we want to relook at anything in light of that. I spoke with my colleague and we said no we felt that what we'd done was right and that was the end of that. **head of department, comprehensive**

Although it was not stated explicitly in very many interviews, a certain number of centres did adjust the CAGs to almost exactly match previous years' performance data before submitting them.

> We looked at the last three years. So our grades that we ended up sending off basically matched what we have had for the past three years. Because we have had consistent grades over the three years and it matched that, so we did take that into consideration as well. **teacher/tutor, academy**

When adjustments were made by SLT they were usually done statistically, based on lowering the CAGs of some students at the bottom of the final rank order within a grade. We have already seen how earlier in the process the use of previous years' data had led to (sometimes hard) discussions involving the departmental staff

deciding which students should come down a grade to produce the profile of grades the centre had calculated.

3.3.3.2 Reluctance to award very high grades

In a couple of instances there were signs that some centres or departments had been hesitant about submitting top-grade (9 or A*) CAGs, related to the fear that this could trigger statistical standardisation.

> A lot of [the hard conversations] were when I was awarding the higher grades. So when I was awarding 7, 8 and 9s. I only awarded two 9s and they got really rigorously questioned about. **teacher/tutor, comprehensive**

> When you give predicted grades you don't want to give a grade 9 because you're too nervous, so then actually you have to give somebody a grade 9. It's quite hard as well, although you'd think it would be easy, because it's not, it's quite hard. So we did, and we did talk to each other a lot about it. **head of department, comprehensive**

There were also a lot of fears that the high grades that had been submitted could be moderated down and that this would not be fair to the outstanding student(s) they had. These issues are covered in section 5.1.

3.3.3.3 Acceptance of submitting CAGs higher than grades awarded in previous years

The most common description of the CAGs was that they were going to be higher than previous years, with a number of reasons to justify or support this. Very frequently we heard comments around giving benefit of the doubt for borderline students, and not predicting random underperformance, as is summed up below and was described in detail for individual student CAGs in section 3.2.8.

> If I was comparing them to last year's performance they were slightly more optimistic. […] You're not going to be able to pick up the students who you thought would get it and actually get into the exam and don't quite, you're just not, you can't predict that and why would you, because often that's a little bit random. And the students who are [a] little bit borderline, you know, if staff think yeah on a good day they could get it, then ultimately if I felt that it's within a level of tolerance then I'd put that forward. **head of centre, comprehensive**

This member of a maths department states the effect of benefit of the doubt on overall CAGs very clearly.

> I would say if I'm really hand on heart honest, I think that probably happened with some of mine who I put down as 5s. I put down 5s thinking really they've got a […] really good chance of getting a 5. Let's say a 70% chance of getting a 5, but actually out of those 10 children in my class who I put a 5, 70% chance they get a 5, yeah great, but three of them won't. And so I'll have overegged it by three. And clearly looking at the collective results from the rest of the department, that was the same. **senior leadership team member, academy**

In addition there was also a conscious awareness we heard several times that there would be fewer U grades as there was no way to be able to predict who would fail their exams.

> The only thing that we didn't do really was that we didn't hand out any U grades. That's one thing we didn't do. So we know that we're always going to have, say, one or two failures, one or two lower grades, but this time [none]. **head of department, sixth form college**

Quite frequently reasons for an improvement this year came up in the interviews. Sometimes this was just a general improvement in the centre or this year's students.

> And obviously we had to bear in mind, actually our cohort this year are brighter than our cohort last year and things like that. **head of department, independent**

> We heard a discussion about the numbers of A*s we should have had and well you should give the same sort of proportion as previous years, well this group have got a higher ability so you can't just do that. They're already at the higher end. They are more likely to get the A*s so I haven't given the same proportion that got A*s last year. **head of department, comprehensive**

A continuing trajectory of centre improvement was sometimes factored in, despite the guidance that this would not be part of the statistical standardisation model.

> Our data shows a slight upward trend, because that's what's been happening over the past three years because we've been able to secure better staff over the last three years, we've made quite a

number of changes in the last few years. **deputy head of centre, comprehensive**

Other times very specific changes were noted to justify higher grades.

But in GCSE we definitely saw an improvement by 8%. But given all the training we've done with teachers and the stuff we did with students, I felt that was fair. And so we agreed on the percentage in between that with senior management and said 'OK, let's see what comes out'. But I'm sure you hear that from others too that GCSE maths has undergone so many things, centre of excellence and all that kind of stuff, we've been so involved. You have to give that some credit, because if you don't then why did we do it all? Well we did it anyway. **head of department, further education establishment**

It's only the second year we've been doing that [course]. So we've learnt from last year's mistakes and got them to a much better place. So, you know, my results would have been better this year and I'm just hoping that it's reflected. **head of department, academy**

Further maths A level […] it was something like 65% A*/35% A, and I said 'in previous years we've had Bs and therefore why are there no Bs?', and he said 'well because this year we've done a better job at persuading boys who are taking maths and further maths if they were struggling, to not take further maths, to drop out, to just focus on trying to get an A* in single maths and they've done that more'. And so all of the boys that have made it this far through the process are the very best ones and therefore there's a statistical reason why it would be different. **deputy head of centre, independent**

And again, a different subject from the same centre.

We appointed a new head of design and technology this year. […] He was confident that this year we're going to do a lot better in their results than the last couple of years' results, and based on the products that they had produced, the physical products, that already seemed to be the case. So we had a long discussion and I agreed that we would put in a much better set of DT GCSE results this year than previous years. **deputy head of centre, independent**

3.3.3.4 Over-optimistic CAGs wouldn't matter

Although most centres used previous years' results to internally moderate their CAGs to some extent, sometimes we also saw an awareness that perhaps it didn't matter, that you could submit optimistic grades and at the very worst they would just be adjusted through statistical standardisation in line with everyone else.

> I kept saying look, if it's too [generous], the exam board will adjust it, that's what […] their job is. This is our prediction, let them adjust it because they will be using their statistics. It's all a done deal probably already, just let them do it. **teacher/tutor, further education establishment**

> I would have thought the schools would err on the trying to be over optimistic than under optimistic. Because you'd be horrified if on results day […] if the statistical model said well you underestimated we would be horrified for that to happen. Much better for you to feel like you thought better of your students. **teacher/tutor, academy**

One respondent reflected on some potential gaming of the system, and was equivocal as to whether the internal moderation or statistical standardisation would fix this.

> They're basically cheating the system at department level within our school. But the checks and balances were catching that out. So it sort of worked but yet you're still getting your 12% inflation of grades. So it's a wider problem than that. So on an individual basis within our school I think there was, it was pretty fair but I am hearing stories from other schools and institutions of, you know, the school's just trying to take the excuse to inflate grades and things like that, and allowing the standardisation process to knock it back, which is what we used to do in coursework and things like this. **teacher/tutor, academy**

3.4 The Vocational and Technical Qualifications centre judgement process

While many of the findings in previous sections apply to the process of producing CAGs (and where appropriate, rank orders) for VTQ qualifications, there were some points to highlight that were specific to VTQ that we describe in this section.

The reassurance of having genuine verifiable evidence was commented on in many interviews as giving high confidence in the submitted judgements.

> It was different because the BTEC is, is it 70% coursework? I forgot off the top of my head, might only be 60. But basically the time of the year we're at, we had all but completed the coursework. So we had really solid secure evidence of where the students were headed. So no, it was pretty straightforward. **head of department, academy**

> Vocational is easier because you keep chipping away at work and its ongoing formative assessment. Exams you're faced with mock exams and classroom notes and workbooks and things like that and that's not an exam. **head of department, further education establishment**

Because of the strong evidence available, some centres were fairly confident that their judgements would stand and not be adjusted, perhaps more so than they would have been for exam-based GQs.

> I'm not so sure that ours will be changed probably as much, because of the fact that it's based more on stuff that's already been achieved. Whereas I suppose with the GCSEs and the A-levels it is very much based on the exams, whereas for us it isn't. **head of department, further education establishment**

For the same reason, one teacher was particularly bullish about their grades for a construction qualification (it should be noted that they reported they were a centre with a proven external quality assurance record) and was anticipating a fight if the AO decided to statistically moderate their results down.

> If I say that's a distinction brick wall but it's really no good, then they'd say well that's not, I get that then, I understand it, therefore […] we're going to knock that child down to a merit from a distinction, and I have to accept that, and I'll end up agreeing with it because it's fair comment. But if they're just downgrading assessment grades because they're there, they're a number on a graph, that's totally unfair, and it's actually morally wrong. It is definitely morally wrong, and I would definitely challenge it. **teacher/tutor, academy**

As well as the strong evidence base giving centres confidence in their submitted grades, external quality assurance by the AOs was also viewed positively.

> We have been asked to supply additional stuff to different awarding organisations and that's not a problem at all. We've been asked to provide a rationale for some of our provision and that's not a problem at all. I think it's important. It's not a negative thing, it's a positive thing I think, robustness in the process. It's just important that we had the

> confidence that we knew we had that stuff here, so that if they were going to challenge or question a grade that was calculated, we had the confidence that yes we had the mock exam results or the sample work. […] We knew that we had the stuff here and we had that confidence. That's really important. And we did, so really positive. **senior leadership team member, further education establishment**

The same senior leader also mused as to whether they had actually belaboured the whole process a little, because they were effectively just carrying on with their normal awarding process:

> In some cases we were asking them to estimate for things that they would have made their own judgements for anyway. So it was adding a bit of extra robustness to the process through a similar meeting structure with a couple of managers, the teacher, managers and an external to quality assure. But actually in retrospect, I wonder if we made too much of it, because the teachers and those managers were used to making assessment decisions anyway. Yes, they would have been submitted and reviewed by the external, by the awarding organisation but […] they had a strong track record, maybe we made too much of it. **senior leadership team member, further education establishment**

However, there were some issues that were faced for VTQs. Many interviewees noted that limited time caused problems for many centres.

> But I think the deadlines that were given were incredibly hard and they were very short. […] I think [the AO] gave us like a week or something. […] That pressure was really quite difficult. […] There were a lot of people stressing at that time trying to get everything submitted and if they didn't understand the process then there were a few arguments between management and staff, not me but other staff in department meetings about how they didn't agree with the deadlines and they didn't understand and things like that. **teacher/tutor, further education establishment**

But the actual requirements of the submissions for VTQ also created some problems for centres. Some AOs were reported as asking for a great deal of evidence to support the judgements, and in some instances centres reported that inabilities to meet these requirements would have stopped learners receiving grades and therefore qualifications.

> The evidence being requested for VTQ calculation is onerous. So I have been asked for one of my VTQs for 24 separate developed pieces of evidence to ratify the calculation […] so I won't be able to

> give them that until they come back in August and as a result those learners will not get their results as expected. **deputy head of centre, further education establishment**

> The GCSE process I think was a lot easier than the functional skills one, because the functional skills was very much evidence based. So I think there were teachers, not myself, but those who didn't have the evidence of the work that the students had done were compelled to predict fails, which I did think was quite unfortunate. **teacher/tutor, further education establishment**

In order to make sure that their students could receive grades, one centre was chasing students for their work in order to assure themselves that they had sufficient evidence to answer any AO queries.

> We had some students who had to deliver some work to us. So we felt confident that there was, that the grade that that member of staff was generating was based on evidence. I had to have evidence. You know I must have used that word numerous times. **deputy head of centre, comprehensive**

The modular nature of the qualifications also made the task more complicated than it was for GQs. Some AOs wanted an overall qualification grade, as well as individual component/unit grades, and making sure that grades for students for individual components were consistent with that for the whole qualification was not straightforward.

> [The AO wanted] each module individually for each student. So yeah, so that's why we needed these 'calculators' that were designed by one of the management, because [...] they wanted an overall grade, but then we needed to submit the module grades as well. [...] Some of our modules have up to five assignments in, so we had to then make sure everything was marked and then submitted for each, then the grade for each assignment can be added up together. So yeah, it was just a lot of work. **teacher/tutor, further education establishment**

Sometimes more coordination was required between teachers than would have been the case within GQ submissions because different units in the same qualification were taught by different teachers, but this did not always create problems.

> Actually there was another teacher teaching the coursework unit and then we did an over-ranking. So I kind of passed on what I thought the rank would be for the exam component and then she did the

> same and in all honesty they were really similar. There was the odd kid here or there that we had to have a bit of a discussion about but yeah it was OK. **teacher/tutor, comprehensive**

For others the different stages students had reached in the course made things more complicated than in GQ.

> They're so complex and the fact that children who were in year 10 in the last academic year, but were also entering their units ready for next year, how they had to be part of the rank as well, that was much more complex. […] It was just the things about units was really challenging, that sense of some students have done these units, some will retake these units, some haven't finished this, it was bitty, vocational was bitty, yeah. **senior leadership team member, academy**

The previous quote touched on the issue of what are known as in-flight learners, in other words, those who are mid-way through their course. They needed to receive grades for the units they would normally have completed by the end of that academic year. This was something that was not a part of the GQ submissions. Awarding grades for these units was important in order to ensure that such learners would not have an unmanageable amount of assessment to complete in their next academic year. However, the inclusion of in-flight learners did create complications, particularly with the unit-based rank orders that were sometimes required.

> What was a problem was I had to submit a ranking of the year 10s alongside the year 11s. And that is the thing that's really caused me grief. Because with 15-year-old students they do change hugely between now and the end of year 11, you know, there's a lot of growing up [that] goes on during the summer holidays for year 10. […] So I've submitted coursework based [grades] for year 10, no problem; rankings, huge problem. […] how the 10s and the 11s fitted in together was really hard. Because it's like, you know, sort of judging oranges and pears alongside each other, they're not the same thing. **head of department, academy**

Some felt that the requirement to provide a rank order in some qualifications was not appropriate.

> I think the rank order for [these] qualifications was tokenistic. I think the idea of ranking was obviously put out there as a thing, how are we going to quality assure it. […] We're going to look at your grades from last year and we'll do your standardisation for you and do your ranking for you, because each crop of learners is totally different to your last crop. So you're bound to have spikes, dips in grade profiles

> and just putting an arbitrary oh let's see what you did last year with a totally different group of learners. **head of department, further education establishment**

Having compiled the submissions, in several interviews problems with the system of inputting the data were reported, firstly in terms of the scale of the data entry required.

> I've got to say for the head of engineering that has been a huge amount of basically data input on an absolutely industrial scale. You know, and I think whilst someone at A level they'll debate whether they get a B in biology, you know, our engineering guy was like, he's got 18 different units. **head of centre, university technical college**

The Head of a different university technical college described an issue with the system used by the AO for receiving the submissions.

> Even after we had input that information there was some confusion about what we had input. And I think we had to go back and do some of it again. And that wasn't an error on our part; that was on their part. And in the end it was OK, and we felt reasonably confident, but during the process that was quite stressful. **head of centre, university technical college**

Centre size and the qualifications they offered made a big difference to overall view. Those with a more varied offer, from a number of AOs, reported more difficulties due to the diversity of approaches required, which created some workload issues. However, centres with a more limited offer reported fairly positive views, largely focusing on the abundance of evidence available on which to base their CAG judgements.

3.5 Summary

It was clear from the interviews that teaching staff had predominantly used a wide range of evidence to arrive at their judgements. In our interview sample the design of the process and the broad evidence types to be used were mostly decided by centre management, although there was a wide variety of decision-making by the individual departments. In fewer cases centres left it to individual departments or staff to make their judgements, using their professional judgement and the published guidance to decide how.

Most interviewees reported a mainly data-led approach, which weighed heavily more objective evidence, such as workbook marks, class tests, coursework and,

particularly, mock exams. In some centres, the use of prior years' attainment data also played a large role in determining the grade distribution that teachers were working towards for the CAGs. It was more typical of larger centres for the process to be dominated by these types of data. In smaller centres, it was more typical that the process involved greater reliance on more subjective evidence, such as student attributes, learning trajectory and attitude to learning to create a picture of how individual students would have performed in their exams. Even where this was the case though, data was commonly used to sense check or triangulate the teacher's judgements.

One particular difficulty was the combination of students across classes, and this was a large part of the discussions that took place within departments. Sometimes mock results were used as the final decision-making evidence to rank order very similar students from different classes. However, not all centres had used the same paper in mocks across all classes, and there were also issues with interpreting the results across classes, for example when teaching staff had different marking standards.

Teaching staff found it particularly difficult to assign grades where students were borderline across two grades, and some indicated that they assigned fewer U grades too, because they couldn't be certain which of their students would have achieved these grades. In these cases of doubt, centres tended to submit the higher grade. But these decisions were not made lightly.

In general, judgements were made using as much information as possible. Even in cases where the mock exams were the dominant basis of the judgements, some additional evidence was considered around individuals, and adjustments to grades or rank orders were made. There were slight differences in the emphasis placed on existing evidence of achievement versus an extrapolation of their current achievement, reflecting different views of how much progress students could make in the final few months. This may also be a reflection of differing interpretations by centres of the information provided by Ofqual and awarding organisations, which emphasised both making evidence-based judgements, and also what students were most likely to achieve in their final assessments, which may imply some judgement of how much effort individuals may have made in their revision.

One thing that came through strongly was that most centres appeared to have implemented detailed systems for checking judgements that involved multiple people seeing and querying them. The judgements were a shared task within the centre, which was never just determined by one single person, and then submitted to the awarding organisation. Sometimes the grades and rank orders could be adjusted as part of this process without input from the teacher though. Class teacher/tutors expressed some frustration with this and felt that the data driving these adjustments was prioritised over their own judgements and experience with the students. But,

given the logistical difficulties of remote working because of closed centres and the limited time available, it did appear as though in most centres a great deal of effort had gone into implementing a thorough quality assurance process.

The general feeling from the interviewees involved with centre assessments for GQs was that the guidance from Ofqual and the AOs was helpful, and the design of the process felt largely unproblematic. However, this sentiment was less shared for those involved with centre assessments for VTQs. For VTQs the guidance and direction generally came later, and interviewees indicated that there was less clarity about what was expected of them. The inconsistency of methods and requirements for centre assessment between VTQ AOs also caused complications, particularly for centres delivering a variety of qualifications from different organisations. This diversity was perhaps inevitable given the range of very different vocational and technical qualifications available.

4 Other considerations in making judgements

The arrangements put in place to produce centre assessment grades (CAGs) and rank orders were new to teaching staff, which meant that without careful consideration, this process could be vulnerable to producing judgements that did not reflect the students' abilities. The interviewees in particular discussed their experiences and perceptions around bias, fairness and pressure; how these issues might have been problematic, how they mitigated them and the degree to which they believed they achieved this. These issues are explored in detail in this chapter.

4.1 Bias and fairness

This section includes perceptions of the fairness of centre judgements for different types of students, how bias might have occurred when making judgements, and measures taken to prevent or reduce bias and unfairness. Bias was discussed with regards to a number of different student groups and protected characteristics including ethnicity, Special Educational Needs and Disabilities (SEND), disadvantaged students (e.g. those eligible for pupil premium and free school meals), sex, type of learner (e.g. last-minute revisers), and behaviour. To help teaching staff make objective and bias-free judgements, Ofqual published guidance (Ofqual, 2020) that schools reported sharing widely.

4.1.1 Information, training and awareness

When asked whether there was any initial training or information sharing around reducing potential biases and ensuring fairness to students with protected characteristics, interviewees described a range of approaches taken by their centre. Some centres delivered formal training sessions or disseminated information to help teachers identify their own potential unconscious biases and how to take that into account when making judgements.

> We developed our own in-house unconscious bias training, so that UB [unconscious bias] training was developed by our equalities and diversity officer who put something together that all teams undertook, just about identifying UB and what strategies they could use to avoid that. **deputy head of centre, further education establishment**

> We had some school training, like a bias training. They'd got […] some guy who was a psychiatrist, [they] did a video of looking out for bias and just trying to get us to really think about unconscious bias and protected groups of characteristics, like boys, girls, SEN, our

> pupil premium students, just having a look at all those students and making sure you're not disadvantaging them because of a certain characteristic. **teacher/tutor, comprehensive**

Many other centres, however, did not have any formal bias training specifically for the process of making judgements. In some cases it was felt there was not enough time to deliver this type of training, or that guidance around bias was published too late.

> No, we didn't have anything. I think because it was all so rushed in a way as well we didn't really have the time to be doing [bias training]. **head of department, comprehensive**

> No, we didn't. We probably should have done, but I think we hadn't got that advice from the exam boards at that point because we [worked on the CAGs] a bit earlier than when that advice was coming out. **head of department, comprehensive**

However, there were often emails or discussions within centres to promote awareness around the issue of bias, as opposed to formal training. Interviewees commented on having previously received bias training before exams were cancelled. They also mentioned other procedures that were already set up to counteract bias by enhancing awareness more generally within a centre's normal teaching or moderation practices, such as generating internal data reports for groups with protected characteristics.

> Yeah, so prior to doing all of this our head governor sent out a big thing on bias. We had pretty much every other day something coming out from our management team about bias, trying to inform everyone and there was a big drive on it within departments as well. Trying to make sure people were well informed, trying to make sure people were considering it. **head of department, academy**

> Every year we do internal reports against those measures [protected characteristics]. We know what we're going to be measured against as a school, and so we're responding really to the lens that's being held up to the school, to interrogate the school by checking internally what we look like. And we know that, yeah, we have regular conversations within the school, with people about how the pupil premium kids are getting on...you know, meeting government agendas that we're looking at all the time. **head of department, academy**

Many interviewees felt that they were already somewhat aware of the potential for bias and were careful to take this into account when making their judgements. In some cases it was felt that their own personal background or experiences were important in helping them understand the issue of bias more thoroughly.

> [Our] teachers are from a variety of backgrounds. I'm a BAME background myself and then the other teachers are all, we're all from different ethnicities, cultures, religions anyway and so having our own personal experiences. I know there's been a lot in the media about centres being biased towards people from different incomes. I'm from a low-income family originally, so that's something that I know that I would then hopefully take into account without actually having to talk about it, to make sure that they weren't disadvantaged at any point. **teacher/tutor, further education establishment**

> I have looked at [bias]. So as I said, I've got, my son is in year 13 and my son has a disability so I have looked at it from that perspective. And I guess I didn't look at, I guess because of my background I feel like I take those things into account quite well anyway [...] I've been quite closely involved in all of the publications that have been released and looked at them all in detail, all the reams and reams of pages. **teacher/tutor, academy**

There were also cases where interviewees commented on awareness of bias promoted by the media. Sometimes they felt this was helpful for highlighting potential unconscious biases.

> I think I was aware of it anyway, because it's been all over the news and all over Twitter and stuff, so it was there in my head anyway. I think if you're a member of staff that didn't go near the news or anything you would have benefited [from bias training] a lot more. But it was useful just to get you thinking and getting your head to be very conscious of being, whether you're being biased or whether you're not being biased. **teacher/tutor, comprehensive**

> I follow a lot of different people on Twitter so I'd read a lot on there about unconscious bias and stuff like that, but I don't think anything came to me from school. But at the same time there was no room for bias because they purely did it on the data. [...] The guy that wrote Boys Don't Try [On Twitter], I think he was quite vocal about some of it and it just made me think. And then I've only got one ethnic minority student in my group and it made me think about him. **teacher/tutor, comprehensive**

However, suggestions in the media that teachers' judgements would automatically be biased was felt to be unfair because it undermined teachers' professionalism.

> But it made me very cross reading that in the papers. […] You know, this idea that as teachers we're going to be so unprofessional as to just go oh I'll give that student a higher grade because I like them, you know, it's not how it works is it? **teacher/tutor, sixth form college**

In defence against the media's suggestions, interviewees felt confident that the risk of bias was mitigated to a large extent by the fact that judgements were rooted in evidence of performance, or due to a rigorous checking process by multiple people as opposed to one single person's judgement (this is explored further in section 4.1.3.1).

One head of centre suggested that specifically because the judgements would be made on evidence of performance alone, bias and fairness training would only really have been effective much earlier, to address the actual performance of particular types of students.

> I think the training side of things would have been quite difficult at that point in time, because what would, you call it training, but what are you actually saying? If there's an issue around student performance for certain characteristics then we need to be addressing that performance and talking about those students on an ongoing basis. [...] We've been really clear with departments that they need to base things in evidence. So where there is an issue around performance of any group then that's the issue that we should be tackling; I think things around bias and any training around that needs to be done on a wider basis before you get to performance issues if you like for an individual student. **head of centre, comprehensive**

4.1.2 Bias mitigation at senior level

Interviewees described a variety of strategies used by centres at a senior level to minimise bias and ensure that judgements were as fair as possible to different groups of students.

4.1.2.1 Comparing data for different groups

Although not all, many interviewees described a data led approach whereby differences in judgements for different groups of students (e.g. by sex, ethnicity, SEND, pupil premium and other characteristics) were compared at various levels (e.g. centre, subject or qualification level), and sometimes compared to the

differences between groups seen in previous years. This type of analysis was often conducted at a senior leadership level, and then used as part of a checking process to challenge teachers where discrepancies were apparent. For example, where a particular group appeared to be underperforming compared to previous years, SLT would do a sense check and make sure teachers could justify the judgements they'd made for these students. It was common for this type of scrutiny to take place. Having multiple people doing cross-checking appeared particularly important for students on the borderline of a grade, where it was not as clear cut.

> We had that information in front of us as well at cohort level, centre level, qualification level, at different levels. So we could [check for differences in] performance in that way as well, if there were any, and we reviewed things at certain points to see if there were any anomalous results coming through for groups in that respect. **senior leadership team member, further education establishment**

> And we also asked [senior leaders] to look at the data across groups, so look at the cohort data and say 'is this what you'd expect, what's sticking out for you here, what are you concerned about, is there something in a particular subject, a particular teacher, a particular cohort, what do you need to go back and test again?' [...] And that often opened up the conversation and that was easier to manage in some ways because it's not just about behaviours, it's about perhaps the teacher not looking clearly enough at evidence. Does that make sense? So we expected our senior leaders to do a lot of that careful work and so supported them with that and they had frameworks and questions to use. **senior leadership team member, academy**

Sometimes an earlier check was done at department level, in this case using an external analysis service.

> So for my role as head of department of applied science, […] we sent all the data to the Alps, which is like basically monitoring software, and they provide very good analysis on basically your white students versus your black Asian ethnic minority students, is there any differences in their performance? [That was] just to stop any bias in your data that you generated. **head of department, sixth form college**

4.1.2.2 Context and individual circumstances

As one head of centre pointed out, although looking at data to highlight over- or under-estimated grades for particular groups is one way of identifying potential bias, it is still difficult to determine what bias is and what a genuine difference in

performance is. Going back to the teachers' knowledge about their individual students helped to disentangle this issue.

> But actually when I look at the work that we've done and the profile of the cohort, it's difficult to know whether that's us then overestimating the boys or it's actually their performance. SEN students, you know, I think staff know their students in the same way as they know any other students. We could take a look at, you know, we do analysis of all our pupil premium students and how they're doing. It's really difficult to separate what you might consider to be bias from actual performance. So we just try to keep the conversation around performance. **head of centre, comprehensive**

In addition to examining data at cohort level, centres also considered different types of contextual information about individual students (for example, their grades across subjects, relationship with teachers, their situation at home) to investigate potential bias. This was sometimes helped by having another person checking, considering the information from a different perspective, or as someone who could view it objectively with less emotional connection. The continuous back and forth discussions between staff members with different viewpoints appeared to help determine the most legitimate judgement for each student.

> I think the hardest thing was that obviously we want the best marks for the students that we can get, but also being honest, and I think that's where having the third person, so having two classes and someone who doesn't have that emotional connection, if you like, I think that's really, really important to have the non-biased opinion. [...] And I think that person who can come in from the outside and say 'hang on but they were like this', and you're like 'yeah that's right, you're right'. And so I think that's where you try and take the emotion out of it. **head of department, comprehensive**

> And our way of trying to make sure it wasn't creeping in was this discussion between different members of staff, because those of us observing the grades they were giving were much more objective about it and if they could be convinced and agreed that they're happy with a grade, then the bias is as controlled as possible. **head of department, academy**

To summarise, one common approach at centre level to counteract bias was a triangulation process between the use of data to look for patterns or anomalies at cohort level, and the use of contextual information available about students at an individual level. These different pieces of information were weighed up through discussions, often with SLT checking for justifications from the class teachers who

knew the students individually. The result would either be that grades/rank orders were kept the same or adjusted based on the new insights.

However, there were a few centres that didn't have such a clear and structured approach, or it may have been that only senior leadership were aware of bias protection processes in these cases.

> I don't know [if anything was done about bias], we weren't specifically told to look at certain individual pupils. That just didn't happen. So I don't know. They might have those sorts of discussions higher up and in the subjects where, English and maths and science that everybody does, but I was not party to anything like that, nothing came down to the options subjects about things like that, so I don't know. **head of department, comprehensive**

> So in terms of bias, no I mean there wasn't anything at all. […] Because the data that we were given to start with was our normal data that we have, when we're doing progress reports we get sent a similar kind of outline and it always highlights those pupil premium in there. So we weren't told specifically to do anything with pupil premium students. I believe, so what we did, we weren't told anything at all but I think the heads of department did look at those kind of things and said things like 'oh we've looked in the past and we have underestimated pupil premium boys', and therefore I think maybe when something happened at a higher level above me they might have looked at changing some of the rankings to take that into consideration. But I don't know. **teacher/tutor, academy**

Sometimes, the main approach was simply to involve more professionals in the discussion, without necessarily looking at data split by groups. For example:

> Yeah, the bias discussion came up a number of times about how we can avoid bias. One of our key things is that nothing was done in isolation in terms of views of an individual teacher. There was a discussion about your groups. So your head of department who may not have taught that group or not known that group, when they were looking through they had a discussion and they probed and had some challenging questions. So, you know, why is that student ranked higher than this student, they shouldn't be. So I gave our heads of department some questions that they had to ask of their team members. **deputy head of centre, comprehensive**

4.1.2.3 Existing procedures

Interviewees highlighted that centres already had their own existing strategies in place to help reduce gaps in performance between different groups. These included

measures to enhance engagement from these groups of students, individualised plans for students with SEN, or adapting how performance is measured within the centre. The benefit of these strategies would already be reflected in the evidence used to make the judgements, and in itself helped promote fairness.

> We've got learning support tutors in each one of our sites, what we call LSMs, and we've also got LSAs, which are learning support advisers. Some of them are purely attached to individual learners from an SEN perspective and we looked at all of that in terms of learners' abilities to engage with the programme, because […] before we even knew of COVID we had extra time for certain learners for certain bits of submissions, etc. [...] That's always the focus of what we do anyway and I guess that is sort of directed via the curriculum we offer in terms of the learners we're offering it to. […] We offer diverse pathways for the learners to engage with. So mainly that they're not disadvantaged specifically from an engagement point of view or an ability to engage because of their background because of the pathways we offer. **head of department, further education establishment**

> All of our SEND students have learning profiles, which make quite clear on one side of A4 […] strategies that work well for them and what they need to achieve and equally some of the things that don't work well for them. So all staff are expected to go with that and where we have students, for example with significant behavioural issues, then we've put something together that's quite similar and often that can be for a student who's got other characteristics as well, we might put them in a more challenging group. **head of centre, comprehensive**

4.1.3 Bias mitigation at class level

There were also discussions around how individual teachers could reduce bias when making their initial judgements within their classes.

4.1.3.1 Evidence-based

Data was relied on heavily again here, particularly where there was ambiguity about a CAG or rank order. Many interviewees felt that by using a data-led approach, teachers were protected against their own personal opinions biasing their judgements, either positively or negatively. The type of data used included tangible examples of students' prior performance, such as grades achieved in mocks or in previous pieces of work throughout the course.

Some interviewees felt confident that their judgements were not biased because of the fact that they looked at the evidence independently, regardless of student characteristics, sometimes even trying not to look at, or talk about, names in case that might influence their decisions. Others acknowledged that as much as they strived to be unbiased, there may always be some element of unconscious bias, but that relying on the evidence and being challenged by someone with a different viewpoint was likely the best way to minimise that bias.

> Well, you sort of like just look at them really, just cold hard data really. I think when I first started looking at the data, I just ignored the student. I just looked at what they'd got and assigned a grade at that point really. It didn't matter what gender they were, what their ethnicity was, I just looked at what they'd got you see, at that point. **head of department, sixth form college**

> When you think of bias, sometimes you, I suppose you have to talk about those students without student name, you know, student X, student Y, student Z, if you look at their data, and that's what I would do sometimes because people would have a very biased opinion, we sometimes have some extremely challenging young people and, you know.[...] It's impossible to mitigate from that because people would have preconceived ideas about that young person, because they may have had a negative experience with them. […] The way that we overcame that was to try and ensure that nothing ever was done in isolation so there was always a quality assurance of that. **deputy head of centre, comprehensive**

While the approach above took characteristics out of the equation in an attempt to encourage unbiased judgements, another approach was to actually highlight the characteristics that should be taken into account on a spreadsheet containing all the information that was available about each student. This was so that teachers were aware of particular characteristics and could focus on ensuring fairness to those students.

> So we had on our spreadsheets, we had whether they had free school meals or had special educational needs or if they had extenuating circumstances such as a bereavement or parental break-up or veterinary exam the day before, so that we had, we tried to get as much information to hand as possible. **head of department, comprehensive**

However, one interviewee felt that being told to focus specifically on some students could even introduce unconscious bias because that in itself differentiated people.

> We were told to look closely [...] in some ways if you're looking at them [at different groups of students such as pupil premium, gender, SEN], isn't that unconsciously biasing you as well? [...] I found that quite a difficult one to get my head around. **head of department, academy**

4.1.3.2 Lack of clarity on what to do with bias information

Some interviewees felt they needed more guidance on how to take into account different student characteristics in their judgements and what evidence they should use. While they were told to carefully examine different groups and be mindful of bias, they didn't know what to actually do with that information.

> I got the idea [factors such as pupil premium, gender] are a focus and look at them and look at your SEN students. But I wasn't quite sure what that meant, do you know what I mean? [...] what am I supposed to do with that information? [...] But we were told make sure you have a good look at these different groups of students. OK, I'm having a look at them, but what do I do with that look at them? [...] I'm aware of it and I know all of this and we've put things in place in school for these students. But how is that affecting the grade that I give them? **head of department, academy**

> I would have appreciated more guidance about special educational needs and access arrangements because what did that mean? So take that one child who'd never had the opportunity to use them, but was going to use them in the summer. What would that have meant for that child? So as a SENCo I didn't really, I felt like I needed more guidance to be able to give to the staff. **SENCo, comprehensive**

4.1.4 Disentangling performance from student behaviour and attributes

There was recognition of the potential for factors other than academic ability, such as behaviour, attitudes to learning, and other student attributes, to influence teachers' judgements. Using an evidence-based approach, as described above, was helpful for teachers trying to avoid being biased by these factors. However, some didn't find it easy to disentangle behavioural attributes from judging how a person would have likely performed in the exam.

There was concern when factors interacted, for example, that students with certain behaviours, commitment profiles or low attendance rates also tended to be those who were disadvantaged for other reasons. In particular, there were concerns that

the rank ordering process and cases of students on the borderline of two grades may be where behaviours and attendance differentiated students, and where bias might therefore have crept in.

4.1.4.1 Low attendance

Interviewees felt it was difficult to ensure fairness to students whose attendance was low, because it meant they'd spent less time with the student and had less evidence of their ability.

> So I think it's difficult when you've got somebody with a low attendance. I think the people with poor attendance were much more difficult to grade than, you know, ethnic groups, disadvantaged in terms of income. I don't think those groups are so difficult to be fair. But I think people with poor attendance who actually may fall into those groups, [that becomes difficult]. **teacher/tutor, selective**

> It's just on those really borderline ones, or like I say the ones who have got very low attendance. But I don't know, maybe there's been a health reason, it's those ones who you just, it's really hard to decide what to do there. **head of department, sixth form college**

4.1.4.2 Relationship with teachers, attitude and behaviour

There were many comments about how students' attitude to work and behaviour were intertwined with their academic performance. If a student was known to be hard working, engaged, and committed, then inevitably that would often be reflected in higher grades on coursework and mocks.

However, there are also some cases of students who despite working extremely hard do not achieve high grades, in contrast to their counterparts who manage to achieve high grades with little effort. This made it difficult for teachers to disentangle behaviour from predicted performance. In some cases a borderline student known to be hard working would be more likely given the benefit of the doubt, than a student who appeared to lack commitment. These types of cases were often where the checking process by SLT, as described in section 4.1.2. would come into play to ensure grades were rooted in evidence and hard data.

> You get a child who skips school a few times a week and comes in, just so blasé and bangs out a distinction. It's unfair but it happens. And that work they've produced without my guidance allows me to distinguish yeah that's a child who can pull it out of the hat, and therefore would do in the exam. **teacher/tutor, academy**

> It is difficult because, yes, some students would turn up on the day and walk out with an A or a 9 or whatever the grade is now; whereas actually when they're in a lesson they may not necessarily be showing that level of engagement, so that's difficult, that is difficult. We don't have a huge amount of behaviour problems, but there are some or there were some in year 11 that their behaviour had been challenging and they had a real lack of engagement and to try and predict their grades was really difficult. **head of department, university technical college**

There was particular concern for cases where a protected characteristic interacted with another factor, i.e. in this case behaviour.

> I've only got one ethnic minority student in my group and it made me think about him, but I was really sad because he's really bright, but he's really lazy so he never does any work, so he didn't do very well in his mock so he's three grades below his target grade and I really wanted to be that person that'd be like no, I'm not being biased, he just does no work, he literally, he isn't going to do very well because he never does lift a finger. **teacher/tutor, comprehensive**

Student personality and relationship with teaching staff was also mentioned as something they had to be mindful of. In many cases they relied upon data and evidence to ensure they were not over-judging students they liked.

> And then also when I have to remind people [when they say] 'but she's a lovely girl'. That's not enough. To say that she's a lovely girl is not enough. Has she got the academic capacity to be able to achieve the grade that you're saying? Do you see what I mean? **head of department, further education establishment**

> For me the nature of my subject and the small class sizes, they are one of my pupils, and so I'm very, very aware of them wanting to do well, and the sort of relationship that being music teachers and working with musicians it is less of a behind a desk subject. [...] You're working with an individual and a personality, it's that wider interaction with them, so very [difficult] and that's why we felt it was safest to use the raw data. **teacher/tutor, independent**

Finally, students with mental health problems were thought to be the most difficult to ensure fairness to by one interviewee. This was because these students could perform poorly on the day of an exam due to their anxiety symptoms.

In truth, the students with depression, anxiety and things like that, they were the most difficult, because you're like well you might have got an A* or you might have got a U. [...] You're looking at students here where you literally have no way of knowing whether they would have got an A* or a U, or anything in between. They could have been anywhere. You literally might as well take a dice, say, six is an A* or one is a U and throw your dice and whatever it comes at you're like OK fine that's what they'll get. We didn't do that obviously. But you would have no way of knowing. Because that's the whole point about human beings, no one knows what they're going to do on the day. So those students, in answer to your question, what did we do with them? Well in the end we basically said well we won't take the anxiety into account much at all, or the depression into account much at all, we'll just work on the basis of more objective evidence rightly or wrongly, I think probably wrongly but you know. The government gave no guidance for such students. **teacher/tutor, independent**

4.1.4.3 Sex differences

Many reflected on sex differences in their classes, mainly their perception that females tend to show a consistent level of effort throughout the year, while males tend to pick up the pace more towards the end of the year, closer to exam time. Although many recognised that this was a generalisation, there were concerns that males could potentially be disadvantaged this year because of this. Some mentioned that they tried to take this into account when making their judgements.

It is a generalisation but a lot of boys pull it out the bag at the end and you would not have had the evidence for it in March. So it all came as a shock. So if we told them in September we are going to be [deciding CAGs], some of them would have given us different work in. And I do feel sorry for them, [...] leaving things to the last minute and that kind of thing, but they didn't know. **head of department, independent**

It is probably the boys, the lazy boys. GCSE more than A level, who just coast along and then pull it together at the end. Yeah, the girls, you know you always have these lovely girls, […] one of these well-behaved girls: […], 'oh she's doing really well and lovely'. And they predict them quite high grades and then they never quite make it in the end because they just work hard all the way through and what they're doing doesn't actually get any better. Whereas you've got the boys who are: muck around a bit and then do a bit of work at the end. Yeah, so and then boys mature different as well. They mature at a different rate to girls and they get to a maturity level towards the end of GCSEs […]. I think from my school, […] which, as I've said [is a] white, middle class school, then I think it [the judgement process] would disadvantage the boys. But I mean I tried to take that [these

> sex differences] into consideration when I was doing the grades.
> **teacher/tutor, academy**

The perceived sex differences above relate closely to the issue of uneven profiles of effort, and how to make judgements for students who would have increased their effort towards the end of their course or relied on last minute revision. This is described in detail in section 6.2.2.

One interviewee commented that there is a positive bias for girls in the engineering job market, therefore this university technical college paid particular attention to a potential gender bias when questioning the justification for CAGs in engineering.

> Now, I think we had to be a little bit careful for positive bias for girls. Because in our world the girls have much better job opportunities, massively better because everyone wants an engineering girl. So a lot of our girls would get 5 job, degree apprenticeship offers, really amazing stuff. So I think we had to be a little bit conscious of that. So I think we bore that in mind when we looked at the [judgements], across the board [...] Now I don't think we explicitly said is it because she's a girl? But I think we said 'well why is that?' **head of centre, university technical college**

4.1.5 Reasonable adjustments, special considerations and access arrangements

The terms "reasonable adjustments", "special consideration" and "access arrangements" can all be used to describe changes to make assessments more accessible for students in different situations. Reasonable adjustments refer to changes to how an assessment is delivered to make them more accessible for disabled students. Special consideration includes other adjustments made to assessments for reasons other than a student's disability, such as illness, injury or bereavement. Access arrangements is a broader term often used within centres to describe any of these adjustments made to assessments.

Different centres took reasonable adjustments and special considerations into account in different ways when making judgements. However, there was a general feeling that more guidance was needed on how teachers should take these factors into consideration. In some centres the teachers simply gave a CAG that they felt the student would have got had they sat the exam, in other words with any of the usual arrangements they would have had in place. They were able to refer to mocks where the students had had these arrangements in place. For special considerations or reasonable adjustments that hadn't been put in place before mocks, see section 4.1.5.3.

We saw earlier in section 3.2 how data spreadsheets were often used by centres and staff to collect together all of the available data for each student. If there were factors that would have contributed towards the necessity for special consideration or reasonable adjustments, this information was sometimes flagged to teachers alongside their performance data on these spreadsheets and therefore became part of their judgements.

4.1.5.1 Considered by SLT

In other centres, the senior leadership team took control of how these extra considerations were handled. Some teachers felt they should have been more involved or did not know how, or if, SLT had made alterations to the CAGs after making these considerations.

> No, as I say, no [considering special circumstances or access arrangements] was all above my head, but I don't think they did and they just purely did it on the day, which as I say I felt there needed to be a bit more tweaking, I think they should have asked me and should have considered those things and then tweaked them accordingly. But they're very panicking about [being] data/evidence driven so they just purely did it on data and evidence. **teacher/tutor, comprehensive**

> So we had told staff not to apply the normal access arrangements. […] We will do that centrally at the end. [...] So we could say look, for example, boys with SEND tend to make, I can't remember what the number we came up with, it was something like 0.4% more of a grade progress between trials and final exams than the main cohort. […] We're going to assume that we can benchmark this pupil against a previous pupil and say we are adding that much extra to the teacher judgement because we need to, because our results don't tally for these pupils with last years' pupils and that seems like something's gone wrong in the process, so we're going to make up for it centrally. **deputy head of centre, independent**

One SENCo felt teachers, not SLT, should be the people taking into account special considerations and reasonable adjustments as these were the people who knew students best. They tried to get this information dispersed to teachers, but it was only taken into account by subject leaders.

> I was communicating to the senior leadership team at school, you must send out [to the teachers] the access [arrangements], here are the children who have access arrangements. And that really has to be flagged up because it's so important to those children. So I did all of that and then the decision at school was that subject teachers weren't given that, subject teachers just graded.[...] So I do have a bit of an

> issue with that, because like I said not all subject leaders know all children. [...] I didn't think it was right that that only went out to subject leaders and not classroom teachers. **SENCo, comprehensive**

One interviewee felt frustrated that although individual students with special considerations were highlighted to teachers, they weren't provided with any contextual details about what the impact on performance might be. Therefore they did not know how much weight to put on the special consideration when making judgements for those students.

> They [SLT] told me that they had things going on in their home life in their mocks and I then tried to get extra clarification about that and that never came, which felt quite frustrating. [...] We were just told something had happened not what the thing was. [...] And in fact one student, it was like it was a completely trivial thing that had supposedly happened and would have no impact at all. So I felt like everybody should have had all the information shared rather than just a flag. Because you can either put too much weight on it or not enough weight on it. **teacher/tutor, academy**

4.1.5.2 SENCo and other support

Across different centres, the knowledge and expertise of SENCos or other professionals, such as the learning support team and pastoral support staff, were used to varying extents to help ensure judgements were fair to those with SEND. Sometimes they were involved in the process of making judgements, either by helping teachers or inputting at SLT level. For example, they were asked for advice, sent reminders about who had access arrangements, and took part in discussions about how to make judgements for those with SEND.

> Yeah the SENCo was, so she's part of SLT anyway. So she was involved and she line manages a number of departments. So she was involved in those sort of meetings. If we had students who we were kind of, if when we were ranking them we had a bit of deliberation between a couple of students and one of them would [be] SEND or a couple of them would SEND and we might have involved the SEN team there. **head of department, university technical college**

> I think teachers always struggle with the concept of access arrangements in terms of the student would have a scribe, the student would have a reader, how does that impact, that was really, really tricky. And again we asked senior leaders to be involved in the conversation and to use the SEND departments because actually we need you to be able to say 'yes, this child should have this and we

> need to be aware of it, don't let the fact that in lessons they can't write very much affect the grade you give them, because actually in the exam they'd have a scribe and a laptop and so on'. So making sure that was part of the equation. **senior leadership team member, academy**

In other centres, however, SENCos were not involved when making judgements. As one interviewee explained, this was because any additional support for SEND students had already been applied within the evidence teachers were basing their judgements on, such as coursework.

> We didn't have any input from any other SEN professionals, or from our SENCo or anything. But we were aware and I read the guidance on obviously if a child normally would be eligible for access arrangements, then we need to factor that in and we need to take that into consideration. I mean obviously with their mock results and things that had already been done, we'd already applied all of those conditions, and all that support had already been put in place and working through their coursework. Which is why we felt more confident basing that on the evidence we've got from that coursework and things, because within their coursework we'd also already applied similar measures. **head of department, independent**

In some centres SENCos were less formally involved in discussions about grading. In one centre, for example, the SENCo's role in the process was only to advise on who had or would have had access arrangements agreed and to answer questions regarding SEND, but not to be involved in judgements about individual students.

> [SENCos were not involved in individual CAGS or rank orders], because they're not subject specialists. We made the decision that they were going to be involved in the strategic, so the award assessment boards, but no, absolutely not. [...] We gave lecturers the opportunity to request specific clarification from that team. [...] In three instances for exceptional candidates we had to use the additional learning support team to gain further information about the work they'd been doing out of class on their subjects, but that was just three instances. **deputy head of centre, further education establishment**

4.1.5.3 Reasonable adjustments or special considerations not agreed before mocks

There were some examples of cases where special considerations or reasonable adjustments had not been agreed or put in place prior to students doing mocks or

coursework, making it difficult to determine how these arrangements would have affected grades. While teachers did their best to reflect this in their judgements, this sometimes resulted in concern that these students may not receive a fair grade.

> For others, especially those who are still going through the dyslexia diagnosis or autism diagnosis, and extra things, particularly extra time for the exams, I think some of them were really disadvantaged. Where we identified them, we worked with the teachers. So I had a list of all the students, and as much as I could I cross referenced, but it was very time consuming to try to be as fair. **head of department, further education establishment**

However, many centres built these considerations into their decision making, for example, by uplifting the mock grades slightly if arrangements were agreed after the mocks, or asking a SENCo for advice on who was likely to have these arrangements agreed.

> And we took into account, […] whether anybody had exam access arrangements in place after the mocks. So if someone following the mocks hadn't finished their paper for example and were then tested for extra time or for a laptop, that was noted in our spreadsheets and so the teachers would know to uplift their grade a bit compared to what they'd actually achieved in the mocks. **head of department, comprehensive**

> So that's why we have the SENCo equivalents in the meeting, because a lot of that conversation would be yes, this learner, you were in the process of applying for X [access arrangement], or X had been applied for, but would they have got it? If they wouldn't have got it, what would that have done to their achievement on that programme? So those were the kind of questions that we asked. **deputy head of centre, further education establishment**

4.1.6 Perceptions of unfairness in use of statistics

There was some uncertainty about how fair to individual students relying heavily on statistics and data was. Firstly, although the centre judgement process was not intended to solve any of the systemic issues in education regarding fairness for some disadvantaged groups, some interviewees felt that although a purely objective, evidence-based approach would be less likely to allow bias, it could not provide a solution to these existing disparities.

> Because if you're being dispassionate, well poor students historically have underperformed. [...] So I think again it comes back to how do

> you make it fair when fairness is different to different teachers, and fairness is different for different students. What is fair basically? If a certain group historically underperforms, is it fair to keep that underperformance, or is it fair to try and address it? Well both are fair, and both are unfair aren't they? **head of department, sixth form college**

Secondly, in some cases the use of data by centres to attempt to align the overall profile of CAGs with the grade profile of previous years was perceived as unfair to individuals.

> I think there was a couple of times they felt that they were too high above the Progress 8, so they went back and started looking. And I think that's maybe when the unfairness and biases might come out when they're trying to keep it at a certain point, because then it's the pupils' grades are then going to get messed around with. **teacher/tutor, comprehensive**

Similarly, interviewees from further education establishments or sixth form colleges expressed concern for their students who were re-sitting English and maths. There was a feeling that the system of predicted performance used by centres felt particularly unfair to these students because they may have already failed to achieve a grade 4 on a number of occasions in previous years, and failing again this year due to a prediction could impact their confidence and motivation to try again. There was concern for resit students who also had a disadvantaged background.

> And then it's hard letting these kids who are predominantly BAME kids failing again. You build them up the whole year, you can do it, and then for the percentage reason, and you end up failing some of them who you think they might have pulled it through. […] And that is for kids from deprived areas a massive thing, because they won't have the drive and the emotional thing to do it again. […] They're turned off forever now. [...] I think I've found it emotionally really hard, working with so many BAME kids and knowing that again they were disadvantaged through a system that dictated the grades to us, rather than what their performance could have been. **head of department, further education establishment**

Dilemmas with deciding which from several very similar ability students should be placed in a lower grade in order to meet statistical predictions were frequently described in the interviews. As well as the perceived unfairness of lowering the grades of some students with very little on which to make that decision, this was also a stressful thing to have to do, something we pick up on in more detail in section 4.2.2.2.

> So basically my line manager was like if we're going to get the value added to zero, which is still a lot better than previous years, we have to basically lower two or three kids' grades basically by one grade. [...] And so we looked at who it should be and we had this big debate over that girl versus the boy that were like the C/D borderline. [...] That boy, he came to every single first year lesson as well as second year lesson, I feel like I can't downgrade him from a C to a D. I probably shouldn't let it affect me, but I was like I know he's not going to be able to get to the university of his choice if he gets a D. […] Whereas I knew the girl had an unconditional offer [from a university] and I was like 'she doesn't need [the grade in my subject], she's going to do [another subject at university]. I did have him ranked higher than her in the end because of the effort he put in, so let's draw the line between a C and a D between those two kids who I felt who very indistinguishable in terms of their previous attainment. So we drew the line so that that girl got a D and the boy got a C. **teacher/tutor, sixth form college**

It should be noted that the concerns described here were contrasted by other views reported in Section 4.1.3.1 that relying on clear objective evidence and data was an effective approach to minimising bias.

4.1.7 Issues for those who could not engage in remote learning

Although the majority of centres did not take into account work done after the 20th March, for VTQs there were comments around the disparities between students who were able to engage in remote learning and those who weren't. For example, some qualifications, such as BTECs, allowed students to re-submit work, thus those that had access to IT at home would have been at an advantage.

> Under the BTEC guidelines, they're allowed resubmission. And there were students who were straight distinction students who'd fallen off the radar completely at the end of March. So they had submitted work pre-lockdown that you had to take into consideration. It had a few corrections to do on it, but to then resubmit and then they disappeared. […] So in some students who had completed stuff it was really useful, but for the students where you knew it wasn't up to their normal standard and they hadn't resubmitted anything. […There was] that student who I told you about, that particular kid who was writing assignments on their phone. And I mean can you imagine trying to write 1,000 words on an iPhone or something. **teacher/tutor, further education establishment**

> But of course like in our college probably about 15% of them don't have access to a laptop, so I'd worry about those students that would be disadvantaged by some of their teachers, because they'll have got a grade below because they weren't able to participate after lockdown. So that's a concern. I didn't base my judgements on their work after closure, but I think a lot of people did. **head of department, further education establishment**

One centre distributed the computers available in the classroom to students' homes in order to overcome this issue.

> Post-lockdown they suddenly didn't have access to the resources. […] So we looked at a lot of the resources that we had that were laying idle in the centres for the most part because of lockdown and then we proactively canvassed all the learners in each one of our provisions where they were disadvantaged in terms of they didn't have access to computers and then we shipped out computers to the learners' home address so that they could continue to engage. **head of department, further education establishment**

There was also concern for students living in chaotic households as they would usually benefit from having revision time in school at the end of the year, away from their busy home environment. There were concerns that judgements may not reflect the positive impact that type of designated revision time would have had on their learning, especially if the judgement was based on a grade achieved in a mock which they had not had designated time in school to prepare for.

> And then my BAME kids, some of them, it's a very deprived borough, so they don't have the computer access; a lot of them only get what they get in the classroom, their work. You know, they cram in the last couple of weeks, because then they stop working, and I think their efforts weren't really represented; many of them could have done better. **head of department, further education establishment**

> But I think that using previous grades, preparation for mock exams in chaotic households or trying to do a mock exam would have caused a problem. And we are seeing, some of the problematic ones are within that group. **teacher/tutor, academy**

4.2 Pressures experienced throughout the process

This section covers discussions around the pressures experienced throughout the judgement process. These include both pressures perceived by the interviewees as well as pressures exerted by various sources on the judgement process.

4.2.1 *Pressure exerted by students and parents*

4.2.1.1 Student and parent contact

An issue raised in several interviews was the experience of students and parents contacting teachers during the process. Given such an unprecedented situation it was no surprise that students and parents would be keen to know more about what was going on, and teaching staff would be a logical place to try to obtain that information. However, in some cases it seemed this contact was made with the goal of exerting some level of pressure to impact the judgements made. This sometimes took the form of providing information about grades required for progression, seemingly with the hopes that this would lead teachers to award these grades.

> And as I said, I know there were teachers who were emailed, they were emailed by pupils saying 'oh Miss I really need to get a grade 8 because I want to get onto medicine one day' or whatever. That was hard for them. **SENCo, comprehensive**

> Actually there were a couple of students […] in my class who contacted me. […They were] really desperate to go to a particular college which required certain grades, and they were saying 'oh I'm really, I'm working really hard, I'm doing this that and the other'. And you're thinking were they trying to influence me? I guess they probably were yeah, and it tugs at the heartstrings doesn't it. **senior leadership team member, academy**

In other cases, a level of pressure was exerted through pleas for consideration of additional evidence or more lenient consideration of certain students' abilities and performance.

> But now we also […] got a lot of parental pressure. [...] And you've got really some quite entertaining emails saying, 'you know, now [student name] did no work in year 10, he did no work in year 11; however he had a damascene conversion and he would have worked really hard and his goldfish died, please can you give him a really high grade'. **head of centre, university technical college**

> And we'd had in the meantime, you know, in the interim lots of nervous pupils and parents saying 'we know that you don't know how this is going to be judged, but please remember that [student name], he did really well in his GCSEs and he had a late burst and please don't just give him his predictions'. **deputy head of centre, independent**

More of the interviews that referred to contact from students and parents reflected parent and student anxieties and concerns rather than the goal of exerting pressure to affect the teacher's judgements. It seemed for some students and parents, the uncertainty around the process and seeking support and guidance for progression prospects was the main focus of the communication with teachers.

> I've not had anybody specifically contact me saying I'm really worried about this, […] but they are concerned about what's going to happen to those grades, I think is actually where the level of concern comes. **head of department, independent**

> And I sought to reassure one or two parents who were concerned about A level prospects for the students that actually whether or not students went onto A levels was probably this year going to be a slightly different process, and what grades they end up with, whilst it would be one piece of information, it wouldn't be the only piece, and we can work our way around things, so no I didn't feel under any undue pressure. **head of department, academy**

A lot of contact from parents or students was made with the aim of learning what the grade for an individual student was, though sometimes this did slightly reflect progression needs, applying a small amount of pressure on the recipient.

> The year 13 students, not so much pressure, but the year 13 students would email me pretty much every day: what's going on, what's happening? What's my grade going to be, what grade have I got in this, what grade have I got in that? **head of department, university technical college**

> We've had a couple of parents email in, and we've just, for A level we just pass them straight to the head of sixth form. So students saying I know you can't really tell me but I've got these two offers from universities which one do you think I should accept, one that's A, A, A or one's that's B, B, B? And you're like I can't answer that! Or else I'd be telling you what grade I'd given you. **head of department, comprehensive**

Or similarly, there were several mentions of parents and students contacting with the aim of learning what evidence was being used and what could be done to improve their grade.

> I think a lot of them emailed, especially when it first came out in the media that it was going to be teacher predicted grades. Literally the next day I must have had about 30 emails just saying can you tell me what my predicted grade is, what can I do to make my grade better? **head of department, further education establishment**

> We had quite a few questions from parents about that, how are you using this data, how much weight are you going to put on it. **deputy head of department, independent**

Regardless of the context of student and parent contact, it was apparent from the vast majority of interviews that teachers did not feel that this contact actually impacted the process or the judgements made.

> I don't think I'd have been unduly influenced by parents asking. We did have quite a few at GCSE and A level but we just passed them on to our senior team to give a reassuring email. **deputy head of department, comprehensive**

> There was quite a bit of discussion with parents on Facebook page that we run, that weren't very happy about it. But that didn't really affect the teachers again. It was kind of some comments removed and taken off, some were replied to, but it didn't really impact on my judgement or anything like that. **teacher/tutor, comprehensive**

4.2.1.2 Effective management of parent/student contact from centre/SLT

The vast majority of interviews suggested there had been little direct contact with parents or students. For those that reported this, it seemed this was a result of effective communication from SLT to parents and students to reduce the contact and potential pressures exerted through it.

For many centres, there was strong communication between the centre, and parents and students. It seemed common that centres would use this communication channel to make it clear that parents and students should not contact teachers, and many commented that this worked effectively.

> So quite early on our head sent out a letter to all parents saying that we would not be discussing any predicted grades with students or parents and that that would remain confidential. And that had been the same. So every member of staff was told that that would be the line that they would have to give that if a student contacted them or a parent contacted them that we weren't allowed to share any information. **head of department, university technical college**

> It was something that the head teacher was, you know, throughout the whole kind of shutdown he was very visible and kept communication lines open regularly, communicated with parents over social media etc. And as part of this process obviously there was a lot of angst amongst parents and students understandably. But I think it was made very clear at every juncture that this is something that can't be discussed. It's something that you can't have an input into, parents. **deputy head of department, comprehensive**

Two centres even commented that they had held online meetings with parents allowing concerns to be addressed directly, for example.

> We agreed to do a webinar, which is the one thing I would never want to do again, a YouTube live event. So the headmaster and I did a live event with questions coming in, so a laptop to one side and parents could ask me questions and I could decide whether I'd read them or not and whether I wanted to answer them! And so we answered a lot of their questions. And that reassured a lot. **deputy head of centre, independent**

In some centres SLT provided staff with generic responses to pass to any parents or students attempting to contact teachers.

> My school were really on that, really on that. They said from the get go 'if you have anything with the students or parents here's the response, copy and paste it, put it out'. And it essentially, it gave them justification and directed anything to them [SLT] if the parents wanted more. So really from a staff perspective we saw very little from outside sources . **head of department, academy**

> So they've [SLT] lifted a lot of pressure off of us and they've given us, we've even had a script sent to us: 'if they say this you do that, if they say this, you say that'. So they're really trying to cover us and make sure we're not put in an awkward position, I suppose, or a compromising position. **teacher/tutor, selective**

In other centres teachers were advised to send all contact to senior management and did not respond to any contact that they received personally.

> But SLT were like 'do not engage, […] do not feel you have to engage in any conversations with them, refer them to us if you think you need to' and that sort of thing. **teacher/tutor, sixth form college**

> Yeah, anything difficult I just forwarded to the vice principal and said 'look this is your job'. And they did, they took that on very well, and dealt with it very well. And so my staff forwarded everything to me. What I could answer I answered; what I couldn't, I moved on. So staff didn't have to engage with anyone, and I think that was very good. **head of department, further education college**

4.2.1.3 Pressure from the prospect of sharing judgements and anticipation of results day

While for the majority, there was little or no contact with parents and students during the process, many interviews suggested a far stronger perceived pressure from the prospect of results day and sharing the centre judgements with parents and students.

> You know, I think a lot of schools are expecting there to be difficulties on the results days this year with parents, with pupils who haven't got the grades that they want, and they will blame us. And so you felt that, you felt very much that it was our responsibility; whereas normally it's not. Normally you'd say well that's how you did in the exam. **head of department, comprehensive**

> So I think we're very aware that on results day the parents will turn to us, if it's not the grade they wanted, that's because we haven't given them that grade. And I think certainly given our context as a school and how vocal some of our parents like to be, we definitely knew that and that was quite hard. **head of department, independent**

This anticipated reaction and pressure, far more so than the genuine contact that some had experienced from parents and students, seemed to have had an effect on the judgements made in the process.

> I think there is a greater anxiety this year from teachers of the comeback, complaints from parents, students, and I think that's probably led to some over-prediction as well. Because if it's an exam it's very easy to say 'oh, well that examiner, we'll have it re-marked',

and you can blame them can't you, you can blame this external force. Whereas it's far harder [with the judgements], as a teacher you've got that crying student or that parent to say well actually 'yeah, I did give you that grade, because I think it's a fair grade'. That's probably played on people's minds as well. **head of department, sixth form college**

We were told that Centre Assessment Grades - we can't discuss them with pupils or parents and you can't release what they are before results day, well, that was great and we thought that's fine, but implicit in that is on results day, everyone's going to go, 'great, what was my centre assessment grade in every subject please?'. And so I know some schools are thinking about just publishing those in full so that effectively we can throw you and the exam boards under the bus and say look I predicted him an A* and nasty old Ofqual standardised it down, which might explain why the grades are 12% higher across the country, because schools would rather it was [you], you guys are the baddies. It would be worse if we gave everyone a B and then it was graded up to an A and they went 'I can't believe you didn't trust my son'. **deputy head of centre, independent**

This anticipation of the lowering of grades by statistical standardisation of the CAGs submitted to awarding organisations led centres to prefer to pass that blame to external sources. This is understandable as it was largely driven by the desire to protect against complaints that parents and students might have, but there was also an interest in maintaining a positive, trusting relationship between students, parents and the teaching staff.

4.2.1.4 Pressure from perceived lack of respect for the teaching profession

A theme that emerged from several interviews was a general perspective that parents, students, the media and the general public lacked respect for the teaching profession. This seemed to add a level of perceived pressure as those producing CAGs and rank orders felt that their judgements and processes would not be respected by the public in general as well as by the parents and students who may complain following the sharing of judgements.

And I don't feel like teachers should have to defend themselves and defend their professional judgements. I think often we're not treated like professionals and that's the danger there. **head of department, independent**

> Yeah, I think that's what will happen […] when we hit results day, is how much those are parents going to trust that professional judgement or not? And given that throughout lockdown we, like I'm sure a lot of schools, have had a lot of parents turning round and saying well you should be doing this. Suddenly everybody's an education expert. I think often a teacher's professional judgement is questioned. **head of department, independent**

While this perception may not have necessarily added to the pressure experienced by teachers, it may to some extent explain the impact that anticipating results day and the sharing of judgements seemed to have on many. It's possible that if teachers felt their judgements were respected they would be less concerned about sharing them with parents and students.

However, we should note that one teacher remarked how positive the concept of centre judgements could be, as teachers' judgements were being valued and respected.

> On the other hand I think it's nice that teachers' professionalism is hopefully going to be looked at, and if all the teachers across the country are being fair and accurate, so I think actually that the teachers do know the students best, then that's also quite nice to think that actually my judgement is being valued and our knowledge of the students is being recognised as accurate. **teacher/tutor, academy**

4.2.2 Pressures from the process itself

4.2.2.1 Practical pressures

Throughout the interviews there were several references to more practical pressures that resulted from the judgement process itself. Some of these related to the time pressure on the process as some teachers were continuing to teach remotely during this time or had additional responsibilities at home.

> I mean the only pressure it added to be honest was the fact that we were doing the grades, and we were still teaching online at the same time. […] I felt the expectation to do this really important grading and also to teach at the same time was a bit much really for those few weeks we were doing it, it was quite tricky. **teacher/tutor, further education establishment**

> Well, it was the timeframe. The timeframe was far too short, because I knew that it would take [the IT team] at least a week to enter that.

> And I was home schooling my child, and trying to do 10 hours, 12 hours of data a day for three weeks. That was really a bit tough to be honest. And going forwards and backwards and keeping your head on it, and being really on it and not losing it within that, I didn't find that easy. **head of department, further education establishment**

One teacher referred to an internal pressure of the timeframe that was determined by the centre, rather than the timeframe determined by awarding organisations.

> I have felt pressure from my department head to do the calculations quickly without the information. I wasn't very comfortable about that at all. So this whole, us meeting this deadline that was much earlier than when they had to be submitted seemed to be very important to school and I didn't feel very comfortable about that. **teacher/tutor, academy**

Another internal pressure felt by a deputy head of centre was the scrutiny they were under from more senior members of staff.

> But I felt because there was such a responsibility in getting it right. […] First of all, it was going to be scrutinised by my principal who is my line manager, and then it was going to be externally scrutinised. And you didn't want to be seen as having been unprofessional, not doing your job properly. **deputy head of centre, further education establishment**

There were also mention of pressures from more senior members of staff in the centre to change CAGs or move grades around (this is discussed in more detail in regard to the internal quality assurance process in section 3.3.2).

> In terms of from school, I think the only time I felt pressure was I think I could pick up that the head or the SLT was saying to subject leaders they needed to push grades down, so hence the email that I had from my subject leader asking me to look at three named children. So I suppose in that respect there was pressure to move kids down. **SENCo, comprehensive**

> So we were using every argument to discuss it and rationalise it and I think that happened three times. I think from memory we had three different meetings over that period, because we were told it was the deadline and then the following week he'd be back saying 'I need a meeting with you tomorrow, we've got to move them, we've got to move them'. […] It wasn't a one off, it wasn't a subtle 'is there anything you can do'? I think he wanted us to change them so that he didn't have to. **teacher/tutor, further education establishment**

A member of SLT in one centre also commented on the pressures of what the outcomes would mean for the centre, which may have added to the pressures experienced by teachers.

> But I don't think we can underestimate internal pressures. We work in a college environment and colleges in particular and schools, because of the way we organise education in our country, are really focused on their outcomes for students and their achievement rates and their attainment rates and their progress measures and their league table positions. But there are pressures there and they can be pressures that the teacher can put on themselves. **senior leadership team member, further education establishment**

However, the most commonly referenced practical pressure was the sheer complexity and volume of work required to complete the judgement process. It seemed many staff felt overworked, as well as mentally and physically exhausted from the process.

> I felt like during those two weeks we were working, all of us were probably working a lot longer and a lot harder than we should have done, because we wanted to do it, obviously. I think our managers were completely aware of that, I mean they were saying just take care, but teachers just do things like that, because they want to be fair to everyone. So it was a hard period. **teacher/tutor, further education establishment**

> But I think from, it was, I probably spent three solid weeks working on, the grades and what went into underpin it and the analysis of it and those sorts of things. So it was very mentally draining. And I did want it to, I'm never a 'that'll do' kind of person. And I think, you know, in these circumstances you definitely can't be anyway. So really it's hard making sure that, you know, you've got it right. You kind of think 'well, have I got those kids in the right order' and that kind of thing. So it was mentally draining. **senior leadership team member, academy**

4.2.2.2 Emotional pressures

One of the most significant pressures experienced that emerged was the emotional burden and the pressure of personal responsibility. It was clear that all involved were aware of the magnitude of the decisions being made and the importance of getting them as accurate as possible. Some of this emotional pressure came from complex moral debates and conflicts due to the impact their decisions may have on an

individual's life and progression. There were several interviews that referred to the difficulty of judging grades for students on the borderline of grade boundaries.

> I think this has been something that I've picked up from other teachers as well, […] for those students who are genuinely borderline between two grades […] how would you pick which half you're saying 'right, no. Actually I'll give you the lower grade'. […] But yeah, I think that's where it's tough isn't it? Because how do you choose which of six borderline students are going to get the lower grade, and which aren't? **head of department, sixth form college**

> What I would say is when they go off and do their exams it's all on them probably, if they've [had] a bad day, and I did feel like for those borderliners where you think it could go either way, that felt like quite a responsibility. Because I'd much rather they go off, get the grade or they don't and that's on them. Whereas, I feel like it's on us now and that's not quite fair. It's not my place to judge a good day or bad day student. **head of department, independent**

More commonly, there were emotional pressures specifically around the borderline of grades 3 and 4 due to the significance of the "pass" grade and what this would mean for the students' life and chances of progression (this was also touched upon in section 3.2.8).

> And what's interesting about that is, I think the assigning of grade 4 in particular was really problematic in English and maths. I don't know if other people say the same thing to you. I think teachers found that incredibly difficult because they realised that's actually deciding whether a student can do their college course or not next year. And if you say they're only going to get a grade 3 you're basically saying 'you've got to retake a year, you might not get into college'. So that's really hard for teachers. Because of course they know the students really, really well and they've been working really hard with them for two years to get them that grade 4 because they know they want to do the catering course, but they still need a grade 4 in English or whatever. **head of department, independent**

Several expressed more general concerns about how their grading decisions would affect students' ability to progress to their chosen college, university or career.

> You feel like you could be holding them back from making progress in a course where they're probably brilliant. […[I mean, I teach motor vehicle and some of these lads are really good with cars, but they're not very good with English. But if you stop them because they can't progress because of their English, I think there's a lot involved in it.

> It's not just whether you pass or fail, it's like putting another year on your training or your life or whatever. **teacher/tutor, further education establishment**

> And then we had some really difficult decisions where we said we are going to stop this pupil going to medical school. It was almost, in each case it was medical school, because in anything else there's a bit of flexibility. But we looked at them and almost out of curiosity I said 'right, if we're going to get this boy who is on A, A, B, if we're to turn his B into an A so he has a fighting chance of getting into medical school because he needs three As, what would have to happen?', and the answer is you'd have to put him from 38th to 33rd in the rank orders. He's clearly not as good as the boys who he would be skipping over and you would have completely incoherent data. **deputy head of centre, independent**

There were also several discussions around the pressure of predicting students a fail grade and how emotionally difficult this was.

> I think I'm more than happy for students to go into an exam and fail it themselves. I know that sounds harsh, I'm not happy for them [to] fail, that's not what I meant, but I'm happy for them to [be] the one that fails or makes themselves fail. I don't like that being me, I don't like saying you're definitely going to fail. That was the hardest bit I think was giving the lower grades knowing that yeah probably they wouldn't have passed, well they wouldn't have passed, but making that judgement of students felt really harsh when they've not actually had their own chance to prove themselves. **teacher/tutor, comprehensive**

One of the most significant pressures that emerged was the personal pressure resulting from the magnitude of the decisions being made. There was a significant theme of personal responsibility throughout the interviews.

> It was stressful, it was really stressful, but I don't know if it was any more stressful than it is getting them ready for an exam. It just felt like a lot more personal stress I suppose, because it felt like you felt you had to get it right, there was so much responsibility to get it right, to be fair to them and to make sure that [...] that you were being as consistent as you could be and that you were doing them justice. **head of department, selective**

> No, I think everyone recognised the gravity of it. Everyone recognised that actually suddenly you've got such a huge impact on these

> people's lives. I mean the opportunity we give them to sit exams is big don't get me wrong, there's a fundamental role of a teacher isn't there? Obviously we're giving them the content, but when it turns round and it's us giving them their grade they're going to carry with them the rest of their life, that is just, it's incomparable to the role we'd normally play. So I think everyone was recognising their role in that, the importance of it. **head of department, academy**

Several interviews referred to sleepless nights and individual students playing on the minds of teachers. There seemed a general theme that the process was emotionally difficult and even if there was clear evidence and an effective process was in place, many still struggled with certain students and felt emotionally burdened by this.

> Yes, a huge responsibility and they didn't take it lightly. And I don't know how many of them would come to the meetings were saying that they'd had a sleepless night, the students in their heads again and again, going through the list again and again each student in turn, thinking 'have I been fair, have I done the right thing?' I know that they didn't take it lightly, so there's a huge responsibility attached to it. **senior leadership team member, further education establishment**

4.2.2.3 Emotional pressures from the context of Covid-19

Some interviews indicated the process was made more complex due to the impact of Covid-19. In several cases this was a result of the isolation caused by lockdown and the requirement to work from home.

> You can only read so many DfE documents and things and you get to a point where you're left feeling, I think for a lot of teachers we felt quite isolated. You're stuck at home not able to go out, not able to meet with the other teachers. You're doing all this and you feel like you've got an awful lot of pressure on you and you're on your own doing it. **head of department, independent**

> So I felt quite freaked out and that was a really challenging period anyway, I've got two young children at home and lots of difficult things were going on, so I was quite stressed. **teacher/tutor, comprehensive**

Understandably, the context of completing the judgement process while also experiencing the challenge of a global pandemic was an additional pressure on teachers who faced a dramatic change to their working routine. It shouldn't be forgotten that this all happened very rapidly, in a few short weeks, from the

awareness of the virus spreading, to the closure of centres and having to work remotely and deliver as much remote learning as possible, to the cancelling of assessments and need to produce CAGs and rank order. While there was no indication in the interviews that these emotional pressures affected the judgement process directly, it is worth considering the number of difficult emotional situations that teachers were facing at this time.

For some though, the impact of lockdown offered one advantage, as it actually reduced some pressure from students.

> But because we were all at home and locked in there was no pressure at all, because they didn't have the access to you in that sense. **teacher/tutor, academy**

> It was lucky that it was a really tight lockdown because I live in the town where the school is, if I'd been out and about I would have seen quite a few of them, fortunately we were all locked away in our houses when this was going on. **teacher/tutor, comprehensive**

4.2.3 Pressures from the media and public opinion

Some interviewees commented on how the coverage of the process by the media added to the pressures they experienced. A common perspective of the media was there was a misunderstanding and oversimplification of the process.

> I think initially, so probably sometime around Easter, when it was announced that second years wouldn't sit exams, […] and I know this isn't what's happening, but the announcement is reported, 'your teachers will pick your grades', […] 'Your grade will be based on your mark'. Really unhelpful stories basically got out into the public domain that were a massive over-simplification of the process. **head of department, sixth form college**

> Yeah, and like I say the messaging doesn't help. They'll have heard it on the news saying 'teachers deciding grades', or 'teachers using mocks', it just ridiculously over-simplifies the process, and suggests that 'yeah, it's just me in a room going I'll give you that, I'll give you that'. **head of department, further education establishment**

One head of department referred to the incorrect terminology used by the media and the negative impact this had on the process and others' understanding of it.

> And I think the terminology of it, […] we're calling them 'centre assessed grades', but if you read any BBC article or Guardian or anything like that it's, they still are calling them 'predicted grades'. [...] And I'm like yeah you do realise it's not predicted grades that are being given out though. Or I've seen people say 'well I was given all the low predicted grades and I beat them all'. And I was like 'yeah but it's not a predicted grade that we are giving, they are totally different'. And I think things like that when the press keep using that phrase, so the kids are turning round and saying well but this is my predicted [grade]. **head of department, independent**

A teacher also commented on their frustrations that the media's focus on teachers being biased was unfounded and unhelpful.

> As a human being, I think there was a lot of pressure groups in the media basically saying […] that teachers are going to be biased no matter what, which I think is slightly unfair, because unless you're in the job you don't know whether someone's going to be biased or not. **teacher/tutor, further education establishment**

In terms of the pressure that this added, the reporting in the media made some making judgements question if they were doing the right thing or whether others were doing the right thing.

> So before, I thought the whole process was fine and fair and straightforward. And initially, I thought we were really good at the way we were doing it. So it's only been since people, you know, the media and that, start making you doubt other people and the integrity of other institutions that makes you then start to worry and become anxious. **teacher/tutor, selective**

There also seemed to be additional concerns about how the media coverage would emotionally impact the students.

> Because when they read news reports about 'oh, kids grades are going to be based on how a school did last year'. That is obviously going to cause anxiety for quite a lot of them, because they're then left worrying, thinking 'well, hang on, why is my grade being based on how that kid did last year?' **head of department, independent**

> But I felt sad for the students as well, because at a time when they were already really nervous and worried about the situation and everything, it didn't help them. I would much rather have had more people vocally sort of going 'we need to trust our teachers, they're

> professionals'. But it's not going to happen. And it's very predictable as to which papers were saying what and all that sort of thing as well. **teacher/tutor, sixth form college**

However, in general it seemed that the majority were unaffected by the media during the process of making their judgements. For many this was because they were too focused on the task itself to pick up on the reporting around it.

> I mean it was there, I mean I've got to be honest with you, because I was concentrating on what I needed to do, I didn't worry too much about it, other than the stuff that was important and relevant to what I was needing to do. **head of department, further education establishment**

> I wonder if after lockdown when there was all the media thing, if they [the students] then started to panic a little bit about things. So I think it caused a little bit of hysteria about it all, but it didn't change what I'd done. **head of department, academy**

Overall, for those that discussed the media, the overarching theme was that the press gave a negative perspective of the process which was unhelpful and frustrating.

> It was really blown up into a big thing […], I think the media made it sound like a really dreadful thing, and in some ways it made it feel more pressured because the media was talking about 'teachers being under so much pressure doing this, and this is so awful, and this is so terrible, and how are they going to get it right?' That you're like 'oh god, is it really that bad?' Whereas really if it had been a bit calmer, I think we would have felt a bit calmer. **head of department, academy**

4.3 Summary

The interviewees discussed several potential sources by which the centre assessment process could produce judgements that were not a reflection of the abilities of different types of students. Interviewees mentioned a number of ways in which they managed the impact of bias and fairness. In many cases, collaboration with other members of staff (such as SLT and SENCos) and continuous discussion about each student's context and the grade and ranking they were given, were perceived as useful strategies.

Centres also often used a data-driven approach to check for parity between different groups of students. There were some teachers we spoke to though who felt that where the onus was on them to minimise negative impacts of any bias or unfairness, more guidance would have been helpful to clarify how to put this into practice and how to reflect this in the CAGs and rank orders.

Discussions further highlighted a range of sources of pressure experienced by those in the process of making judgements. On the whole, the interviewees reported that there was minimal contact with parents and students that aimed to influence the judgements made regarding grades and rankings. In most cases, these types of communications were diverted away from the teachers making the judgements and managed by SLT. The teachers felt very well supported in this respect.

The main source of pressure to influence grades came from the anticipation that parents and students would be disappointed or unhappy on results day. The teachers often talked about personal responsibility for the grades the students would be awarded. They reported experiencing a lot of stress and anxiety around what the grades would mean for students, and whether there would be negative consequences for their progression. This was particularly difficult for students who were less secure within their grade. For some, they reported that these issues may have resulted in some generosity in the grades they produced.

In some discussions, interviewees also reported experiencing pressure within the centre. This was particularly related to requirements from SLT to change their original judgements to be in line with more data-driven predictions. Interviewees also commented on how the narrative in the media about the judgement process had been unhelpful and as a result felt increased pressure to 'get their CAGs right', to protect integrity in the teaching profession.

5 Beyond the centre judgement process

This chapter explores aspects of the assessment process after centres had submitted their CAGs and rank orders to the exam boards. In particular, this section addresses discussions teachers had around the exam boards' standardisation process that Ofqual put in place, views on the provision of the autumn exam series, and any worries regarding the subsequent academic year.

5.1 Standardisation process

Those interviewed were aware that for GQ and some VTQs there would be a statistical standardisation process to produce a calculated grade that students would receive, and that this was to be applied by the exam boards after centres had submitted their judgements. Some of the comments in this section reflect slight misreading of how the final model would work. Nonetheless, we report these as they reflect concerns that were real at this moment in time.

At the point of interview, the outcomes of the standardisation were not known, and the decision to switch to awarding the higher of the CAG or calculated grade was still some weeks away. It should be noted though that a small number of interviews took place after the standardisation process had been undertaken in the Scottish assessment system and results issued.

As we have described in sections 3.2.8 and 3.3.3, there was a common awareness that some of their submitted CAGs were potentially a little high, particularly for borderline students. Some interviewees did anticipate that CAGs could be downgraded.

> There might be, I think there will be a couple that will shift maybe down, maybe up. I think it's more than likely for people to move down than to move up if you know what I mean, in terms of my cohort. But yeah, I'd be disappointed if there was too much movement I think. **teacher/tutor, sixth form college**

> And that was like I said before, also my thinking where perhaps I gave someone the benefit of the doubt but put them at the bottom of the rank. So I might have given them an 8, they might slip down to a 7, but that was my reasoning with that process. So I'm expecting that that might happen in some cases. **head of department, selective**

> I think that the grades will get adjusted downwards. I'm talking nationally here. I think our school's [grades] will go down a little. I

think our 5s will go down a little. **senior leadership team Member, academy**

Not all teachers thought that the standardisation would impact upon their grades. Many expressed that they did not expect their grades to be changed at all. This was most common in centres which had matched their CAGs to patterns in their historical data so that standardisation would not be necessary.

> I will be honest, if there was a huge amount of adjustment I will be surprised. Purely because I think actually the centre assessment grades that we have submitted, statistically in terms of the grade distribution, it's so similar to what we, to our historical evidence that if there was to be a huge adjustment down let's say, I think as a school we would be questioning that. **head of department, independent**

> I'm hoping for none [no adjustments to grades following standardisation]. Like I say I'm fairly confident with it based on the various stress tests that we did with it. I think as an institution that what we've done will fit the modelling. **teacher/tutor, sixth form college**

Other people felt that while the CAGs produced in their own centre were trustworthy, CAGs produced by other centres were likely to be inflated, and the result of judgements made through a centre-based process that was not as thorough as their own. For some, there were concerns that the actions of other centres would impact their own grades through national standardisation.

> My biggest thing is, and this sounds awful but I think it needs to be said, is trusting that other schools are going to do it right. And I think the media didn't help with that at all, because they were saying 'schools are going to inflate grades', and that can play with your mind a little bit, because you start to think well if all schools are going to inflate their grades, they'll all go down a grade, so should I inflate mine, because everybody's going to do it. **head of department, academy**

> So because other organisations, educational organisations will have inflated their grades shall we say, statistically, nationally. If you put it all together in the pot nationally you couldn't have an extreme rise, I get that. So I think everything, the modelling that they use and moderation process will probably cap us at just a fraction more than we've ever been - which is a shame, because we've been, like other

> people would say we've been working really hard at quality, but it will be what it will be. **deputy head of centre, Further Education**

Many felt that because the standardisation model would not take into account how centres could be on an improving trajectory, these centres or departments within these centres would be disadvantaged. As we saw in section 3.3.3.3, some centres believed they had a sound basis on which to expect improvement such as new management, better teaching or differences between cohorts. They feared that this would not be recognised, and instead CAGs higher than previous years would be reduced by the standardisation model.

> It seems to me that they were so weighted on previous results, and I think that's, well it'll advantage some people, and it'll disadvantage others I guess. The people I feel sorry for as well is the schools that have gone into special measures, and then they've had new staff, new head teachers that have turned a school round, but they'll be so burdened by past results. **head of department, Further Education**

> We're all saying, it's so frustrating because this year, this cohort were really good and we were all on for a really good set of grades which we now don't get to use in our sort of like 'well, the department got 100% pass rate'. So very frustrating. **teacher/tutor, sixth form college**

This teacher objected to the idea that results would be held to previous years, and lock in any existing differences between centres.

> The standardisation model does look like it's going to be unfair to people, where departments are changing over time. Good schools are going to get good results; poor schools are going to get poor results. So there's problems with that. **teacher/tutor, academy**

A few teachers further commented on specific circumstances, and expressed that they were worried that the standardisation process would remove individual students' improvement if this did not fit the usual patterns.

> But in our school we do have those kids that get really good results just in one or two subjects, and that's their livelihood. […] If they're profiling based on prior data, key stage 2, which they said they are, they're looking at the last three years and stuff like that. […] [For] those particular children [who aren't strong across the board], I'm worried that because they are low [attainers in many subjects], that

they will be treated as low [attainers] in everything. **deputy head of centre, academy.**

It was not always clear to teachers how the standardisation would work, particularly in VTQs. But there was a clear view in these cases that grades should not be adjusted outside of the centre without considering the evidence which formed the basis of those judgements.

> I don't know how they're going to standardise it all, I don't know how they're going to verify anything, because I've actually said all my evidence is here. I would have gladly sent it all off, I would have got it all packaged up and sent it all. […] If they've not asked for anything off us, as in proof, they shouldn't be touching our grades. **teacher/tutor, academy**

Contrary to some of the views above, there were many who had positive views about the standardisation process. Some expressed that they were reassured by the standardisation process because they felt that it made the system fairer by reducing potential overinflated grades. Others felt relieved by the reduction in responsibility they had for the final grades.

> So part of me thinks that Scotland probably got it right initially [to apply the standardisation model]. […] I think teachers have a lot of influences and I'm sure that if you told me that it was definitely just my grade and it wasn't going to be moderated, I'm sure I'd have gone higher, because you think that every other school would have done, would have been more biased if we'd been told that there wasn't a moderation process. **head of department, comprehensive**

> There's a certain level of protection for me as a teacher, because I've put the grades in and then the exam boards are going to do something with them, so it takes a little bit of the pressure off me with the grade I felt. […] So I felt there was a nice […] wall between the teacher and the final grade. **head of department, comprehensive**

There was also the tacit recognition that standardisation was a necessary part of the assessment process, with its function largely to correct generous grades because it was important for students to progress onto courses that were suitable for their level of ability.

> But equally, we're not doing them any favours if we let them onto the course if they can't cope with it. So, we'll be just seeking to have really honest detailed conversations with the young people before

they do the A levels in the way in the past we wouldn't have. So that first week back in September is going to be crucial for many students. **head of department, academy**

5.2 Autumn series and progression

Interviewees also often discussed the provision of the autumn exam series and appeals, and the issue of progression to the next stage of education for some students. We report these views in full as they do relate directly to the judgement process as a whole.

Many interviewees felt that the ability to sit the exams in the autumn made the system fairer for those students who were not happy with their grades.

> I thought the announcement of the possibility, if you're not happy with your grade of doing the GCSE exam, I thought that was genius. […] I don't see what else they could have done and it tackles the student who claims that the CAG is not fair because they were going to work and they haven't started working. Because OK great, get the work done, come in, do the exam. And I think that's fair. **head of department, academy**

Of course, there was also recognition that this was an imperfect solution due to the delay in progression this could introduce.

> So all they will see is the importance of the November opportunity to get that grade. In which case it's a shame because they're to not get a result what, maybe until January. […] But I don't right now see how it will really work for young people, because it involves putting whatever they're doing on hold, until January really, waiting for that grade. If that grade was going to allow them to do something that they're now not able to do then they're having to pause. **head of department, academy**

This delay in progression was viewed by several interviewees as particularly serious for students from more disadvantaged backgrounds who may not have the option to take the autumn series examinations, perhaps because they could not financially afford to take another year out if it delayed going to university.

> And there is the option of […] the exams in October. Which again I think is going to disadvantage certain, so you're going to get the ambitious ones, the ones who can take the extra year. Some from more disadvantaged backgrounds will have, might have problems with that...**teacher/tutor, academy**

> And I don't want my students to be downgraded because, especially those vulnerable ones who've got some fantastic offers from [good universities]. You know, they were doing these contextualised offers, don't take that away from the kids. They're ready to go. They have battled much harder than many kids from selective colleges. They have the skills to survive. And they have the network to survive. Don't disadvantage them now. I mean if they get disadvantaged because of this, then the October resit, what is it going to give them, they're not going to university for a year. **head of department, further education establishment**

However, one teacher felt that if universities were flexible with entry then students from disadvantaged backgrounds would have a chance to get into university and use the resits to confirm their places.

> So a lot of my less ambitious, less affluent backgrounds, their objective is to get to university and through this system they will. So in a way it's not so bad. If you just have this system and let the universities sort out the thing when they can actually do formal examinations I think you'll still end up with a fair system in the end. **teacher/tutor, academy**

Some interviewees did express frustration with the autumn exams series though, suggesting that it would be a poor substitute for an appeals system for the grades awarded to students in the summer, and that students would just not be prepared adequately for it.

> And I think to say 'you've got no chance of appeal, don't worry though, you can do another exams season in November', is absolutely outrageous. Because what child in March really will have retained anything, especially those who are the critical borderline ones. So that is heart breaking. So yeah, a bit difficult. […] The reality for these children is they've not had a maths or an English lesson since, well I don't know, was it 19th March, 18th March, something like that. […] But if they had done it as normal, they'd have been absolutely eating, breathing, sleeping maths and English right up to the exam. **SENCo, comprehensive**

> There is the option of doing an autumn exam. And we've said that to everybody, you know, that's kind of been the pacifier for people hasn't it. If you don't think you got the grade you deserve, sit an autumn exam. […] I'll be really interested to see the uptake because they haven't done anything since March. You know, an A grade student in

> March, […] to sit an exam seven months later without that constant preparation from their teachers. […] there's just no way that an A grade student in March is going to produce an A grade in October without seriously devoting a lot of time to it. They're just not going to be exam ready. So I don't know that people would want to particularly bother. **teacher/tutor, selective**

Others considered the practicalities of the autumn exam series. They had concerns about the resources that would be required if there was a high number of students sitting their exams in autumn. Some also questioned whether sitting exams in autumn would be useful.

> I think that's going to be the toughest part, where do we actually do these exams? I think more the logistics more than anything though. Where do you actually do them? Is it actually feasibly possible to do them at the moment? That for me is the toughest part. And then the students have got to prepare as well because, then we talk about every student will have to revise. They may expect further support from the class teachers etc., and ethically I've got to give them that support. If they ask me for it I've got to do it, I can't just leave them. **head of department, sixth form college**

> And that's a massive resource issue for most training providers and colleges, because everybody works around the June exam period and obviously if suddenly you've got another potentially substantial exam period in November […] So that's a worry. We've got things in place to try and advise learners against jumping in too early, because they might not get the result if they do it in November anyway if they're not ready. **head of centre, further education establishment**

There were a few comments on the wider impact that the centre judgement process might have on entry to higher education. One interviewee from a further education establishment felt that students on the 'Access to Higher Education' course were overlooked in the consultations and may be disadvantaged compared to A level students applying to university.

> Throughout the consultation process and throughout the arrangements, there was always questions about why they [Access to Higher Education students] were being treated quite differently and the potential there for disadvantage. [In the] first two weeks of April, for example there was information coming from Ofqual and information coming from universities and Access to Higher Education kept being missed out of the discussion about progression to universities and how to make sure that students weren't disadvantaged. So the discourse was consistently about A level

> students not being disadvantaged and universities giving flexibility to A level students [...] And that discourse never included Access to Higher Education students who are aiming to do the same thing.
> **senior leadership team member, further education establishment**

It should be noted that these Access to Higher Education courses are not regulated by Ofqual, but are instead overseen by QAA. They were therefore not part of the centre judgement process in this research.

5.3 Worries regarding the upcoming school year (2020-21)

Although not one of the topics we wanted to include in our interviews, the issue of the up-coming school year for the next cohort of students did come up frequently. We did not formally code this part of the interview, but here we describe in broad terms the worries interviewees expressed.

Many teachers were worried about the students starting the final year of their qualifications in September 2020, and the level of disruption they had already experienced and would face in the future. There was some discussion about the planning that had already started around the return to school in September and the stress of getting this right. They mentioned the variety of different scenarios they were having to plan for and how difficult it was to predict the future and respond to it. Fears for how the students could possibly catch up on the lost learning in the previous year, and the likelihood of some level of additional lost learning in the coming year were frequently touched upon.

As part of their reflection on how the 2020 centre judgement process had gone, some interviewees also commented on the prospect of having to repeat the process in some form in 2021, how that might work, and what steps they would put in place to support it, such as more standardised (cross-class) testing or banking of coursework or classwork evidence a little earlier in the year. Some respondents did express their confidence that they would be able to repeat the process with an even stronger evidence base, even in cases where they would prefer not to have to.

Some discussion of exams in summer 2021 also occurred, including what changes they would like to see in the tests, such as optionality or a general reduction in the amount of assessment. Overall, the overriding impression was that while the situation had not been ideal for the students for which they had made judgements in summer 2020, those students might have been fortunate compared to the students facing their final year of their qualification in 2021, who were likely to have experienced significantly more learning loss.

5.4 Summary

Discussions about aspects of the assessment process after the centre judgements had been submitted tended to focus on the role of statistical standardisation put in place by the exam boards and some AOs in VTQs, and the degree to which this would result in the reduction of CAGs. Some were confident that their centre assessments honestly reflected their students' capabilities, and believed that their CAGs would not need to be changed as part of the standardisation process. For others, they suspected that there would be at least some reduction to the grades. Those who held this view expressed that this could be likely, especially where they or other centres had submitted grades that were perhaps a little generous. For instance, many recognised that they had given the benefit of the doubt for borderline students.

While many reflected on the impact changes to CAGs would have on individual students, some interviewees held the view that standardisation was a key part of the assessment process. This was driven by perceptions around the necessity to align grading across centres, particularly where they had been some leniency, and to ensure that students were appropriately prepared for whatever they progressed onto. Other interviewees had concerns though that achievements for centres, departments, or individual students with atypical improvement trajectories would not be recognised.

Some interviewees also mentioned the autumn exam series. There were some who felt that this offered a suitable opportunity in which students could prove themselves if they received a grade in the summer that they were not happy with. There were others, however, who were less positive about the autumn series. They had concerns about the degree to which the students would be prepared to take these exams, and felt that they would disadvantage the already disadvantaged, particularly those who had fewer opportunities to undertake learning.

Senior members of staff also considered the logistics of the autumn exam series and questioned how they would accomplish these on a practical level. The interviewees also reflected on the students due to take their exams in 2021. They were particularly worried about what the assessment process would look like and had fears that that cohort would be even more greatly disadvantaged due to the disruption caused by the pandemic.

6 Overall confidence in the centre judgements

This chapter explores issues that emerged from discussion relating to teachers' overarching view of the judgement process. Many of the issues explored in the sections above, such as bias, fairness, the process by which the centre assessments were agreed, and the evidence used to produce the CAGs and rank orders, all contributed towards their final views. These issues are not explored again in detail here, rather, discussions within this theme are generally related to how these issues impacted their confidence in the whole judgement process. The degree of confidence reported was entwined with issues related to reliability and validity of the judgements, within their own and in other centres.

The concepts of reliability and validity were typically explored by participants comparing the centre assessment grading process with the exam assessment process. In general, there were mixed views regarding how reliable and valid the judgements were. The issues and discussions relating to these two concepts are set out in turn.

6.1 Perceived reliability of centre judgements

Reliability here refers to the degree to which the centre judgement process was a trusted means to measure student performance. Nearly all who commented on the issue expressed how they believed the CAGs and rank orders they were responsible for were reliable. For some participants, they were confident in the grades submitted because the process was the same, if not similar, to an existing process they believed was a reliable way to make grade predictions.

> Because my job is to sort of coordinate and track the progress of the key stage 4 year group, I do it throughout the year anyway […] I've always got an eye on how kids are progressing when we've got the unit assessments, and the tracking and monitoring that we're doing. So I'm quite comfortable in making those predictions **senior leadership team member, academy**

This was repeated in quite a few interviews, that centres collect vast amounts of data and routinely do something 'close to CAGs' every year, so giving them high confidence.

> For the A level we assess all our students all the time. We're […] generating data every single week you see to […] help our students progress. So like I say the A level chemistry, we had a lot of data all the way through for the 18 months. We had numerous tests, we had

> numerous mock examinations, and so therefore we had that data to show where the student was at, and what we think they'd get with the idea of the upward trajectory. **head of department, sixth form college**

> I know we started with the teacher grades, because we already put targets and expected grades in. So it wasn't generating any new information because we already had an idea of what we thought they would get. So it started there. **teacher/tutor, comprehensive**

The depth of data analysis available was described by this senior leader, who spoke about the use of software tools or external resources to manage their data, and these tools were frequently referred to in interviews either as a starting point for the judgements, or a way of quality assuring them.

> There's different data analysis […] software out there, […] they're similar things where you basically put your data in, and […] you can then analyse that. […] It would show you whole school results, it would show you Progress 8, it would show you your different groups in terms of boys, girls, upper attaining students, your pupil premium students, etc. It breaks it down every which way you can, but it also it would show you how many grade 8s, how many grade 7s, how many grade 6s. So they'd put that into SISRA first to have a look at what that would show, and then think well actually are we happy with that really. So they'd got that pre-analysis. **senior leadership team member, academy**

This use of normal data capture and analysis was further supported by comments that previous years' predictions about what grades students would achieve had been accurate.

> I feel confident that they are as close a reflection as we could have got. I'm also confident because we've done a lot of work. […] and last year our predicted grades were pretty spot on. So I feel fairly confident in terms of the CAGs that we submitted. **head of centre, university technical college**

> So the fact that we've used a load of data and it's supported, and we've generally been within 3% or 4% in previous years, when they look at our data, the centre assessment grades, really they would only have moved let's say 3% or 4% anyway, so we're not going to have grade inflation. So it's likely then that [standardisation will say] that's exactly what we expect them to get. **head of department, academy**

Some commented on how they felt that the evidence they based their judgements on was high quality, and had sufficient breadth and depth.

> I think because we were already quite well set up, you know, we'd got our data, I think as well because our data came from mock exams that were in the new year it was quite reliable. **teacher/tutor, selective**

> So I suppose we were lucky that we had some really solid agreed standardised data that [my colleague] and I had worked on together, **head of department, selective**

The collaboration with several different members of staff who quality assured their rankings, grades and justifications for them, further gave them confidence that the right judgements had been made.

> I have no idea if these are the grades they would have got, they are my best, our best estimate, and there's four teachers involved in my chemistry team in putting this information together, me and three other teachers. So I feel really confident that we've done our absolute best. **head of department, comprehensive**

> And I do feel pretty confident because once we had been through the process as a team, you know, for example the English team talking through it and then feeling confident to submit their rank ordering and their centre-assessed grades, well then that was quality assured and that conversation went backwards and forward two or three times before we were totally confident of the head of department signing it off and saying 'this is it now'. **deputy head of department, comprehensive**

Some interviewees also reflected on how their experience in the teaching profession, and as examiners, enabled them to understand how students' skills and abilities mapped onto grades.

> At the time that I was doing the CAGs I would normally be flat out marking exams scripts, which are then standardised. So I know how the system works. So yeah, that's why I say I was quite confident in what I was doing with it. […] Because we'd both been in that position, we felt quite confident that we were there or thereabouts when it came to marking. And then we could have those discussions about whether we felt we were being too high or being too low. But we both

had a bit of an awareness of what the standard was having been on the other side of it. **head of department, selective**

Given the various exam board positions I hold and the fact that I'm quite old and I've been teaching for quite a long time, I didn't particularly find this very difficult. […] The mock exam that I set for my GCSE students was, well there were two papers and both of them I was heavily involved with marking last summer. So when I set those exams so far as I was concerned if they'd sat that exam on that day that is absolutely what they would have achieved. So I feel like personally I was in quite a privileged position to feel that confident and to be relatively sure that that would have been the outcome at the time. **teacher/tutor, comprehensive**

A few respondents touched on their centre having a good history of coursework moderation, increasing their confidence in their judgements.

I'm very sure of the grades that I put in. I've been a head of department for over 20 years and I've done a lot of, we do moderation in drama anyway and my moderated grades had never been changed. So I was very sure of it. **head of centre, academy**

We were quite confident, our coursework marks have never been adjusted. We have a member of staff within our department who is an examiner for the exam board. And we've done quite a lot of work this year, actually around the mocks, of adjusting our internal marking to fit with the marks with how the exam board seem to mark. **head of department, independent**

Many shared the view that there should be trust that the centre judgements would be reliable because of teachers' knowledge of their students and their integrity.

I think that teachers should be trusted a lot more in terms of what they know and how much they know about the students. So I think that what will come out of this is that actually teachers do have a really good sense of what's going on and that it reflects the fact that students' learning is really learning. **head of department, selective**

I think it is possible to do this well, through moderated activity [and] quality conversation that we've done, I think it's absolutely possible to do this with accuracy and to trust teachers. Teachers do know their students better than anybody else. **senior leadership team member, academy**

A few participants further expressed confidence in the centre assessments, indicating that they hoped the opportunity to use teacher assessments this year would spark a change in the assessment system in the future: to enable more teacher judgement in the overall grades that students receive.

> Part of me wishes that this would be the start of a change in the British system, where we're not going to rely so heavily on these exams. Because I think that there's so much to be gained from a broader assessment of individuals' knowledge of your subject that's more than just an exam. **teacher/tutor, sixth form college**

> I think we do need to look at doing things in different ways and this has been an opportunity to do that. And so I think we do need the debate to be focused on the curriculum of the future, the education of the future, the system of the future and this was an opportunity to maybe do things differently. And I do honestly believe that centre assessment, centre calculation is a positive thing, allow teachers to calculate, why not? **senior leadership team member, further education establishment**

A few participants also drew positives from the process relating to professional development. For example:

> It was a great process internally, also to see about your staff, where their strengths are. I mean, yeah, we ended up with a lot of interesting conversations and a new standardised plan for next year. So it had some positive outcomes. **head of department, further education establishment**

Not all interviewees were happy with the judgement process though. A few expressed that they had limitations in their confidence in the process. For some participants this was because they, at least in part, found it difficult to rank the students. Examples of this include simply where the participant didn't have confidence in the ranking process at their centre, and in another example a teacher commented on how it was difficult to compare students who had different approaches to learning.

> They've now said they're introducing these statistical algorithms which will help to maybe sort problems out, but I don't think they will, because they still have to use the ranking system that we sent in, and I don't think our ranking system is accurate. **teacher/tutor, independent**

134

> I would say, in terms of the ranking, certainly at the top end I had a couple of students who I honestly would have found it really difficult to split between, because they were very similar ability-wise, but actually just a very different approach to how they learnt. So one of them was very quiet and studious, the other one was very outgoing, very chatty, wanting to answer questions in class. And it just seemed really difficult to establish between them on the basis of being so different, which as to where to put them in the rank, it really was. **teacher/tutor Comprehensive**

Although we have seen throughout these interviews how teaching staff felt that they usually had plenty of reliable evidence on which to base their judgements, it is important to note that a few in our sample commented on how, at least for some students, the evidence they used to arrive at their judgements wasn't as reliable or complete as they would have liked.

> And the other thing that didn't happen this year, which normally would [have], is the year 13: the work they did in year 12 would normally have been [externally] verified by the standard verifier about February. Year 13 were just about to start that process. So I'm confident having done it for a few years now for year 13s that I can grade something accurately. But I do like it verified as well. So it's been internally verified within the centre but, you know, it's not as secure as I'd like it to be. **head of department, academy**

> Am I confident in my numbers? Well I'm not confident in the ones where I thought I had missing information. And so my A level class, I taught them for two years and I felt I did have all the information I could have had. I feel, well I feel my grades will definitely be right within a grade. But it's still not very accurate is it - not really. **teacher/tutor, academy**

In general, there was dissatisfaction amongst teachers where they were aware or suspected that SLT had made changes to the CAGs the class teaching staff had worked on before they submitted them. We have touched on this issue in section 5.1, but we just repeat that some interviewees felt that these adjustments could make the CAGs unfair for some students. For example:

> So I ended up [giving] what [grade] I felt that they were going to get, […], and then we didn't hear anything and I kept emailing my head of department saying 'have we got any news on things?' […] The head still felt that as a school, compared to how we'd done in the past it [the grade distribution] was still too inflated and he, as far as I'm

aware, on his own in consultation with SLT, regraded everything. Now I don't know what my class looks like now. **teacher/tutor, comprehensive**

Many interviewees had further distrust in the way in which some other centres might have made their judgements. This was largely fuelled by the perception that because grades are high stakes, and have implications for issues such as student progress, performance tables and teacher incentives, some centres will have overestimated them.

In other schools, maybe there was a conflict of interest where you've got performance related pay. **head of department, independent school**

If you've got a system that trusts teachers to rank and grade their students, but then you're giving them incentives to massage those figures because of league tables, or because of how they will look, those two things don't work, because of course they won't, well of course some won't use integrity because they've got a reason to cheat the system. **head of department, sixth form college**

Some participants also cast doubt on the robustness of the judgements underpinning the whole process in general.

You know, teachers have off days, teachers do have bias you know, I don't think this is a good process at all. **teacher/tutor, academy**

I think the other thing that I'd say, I don't know, just something else that would concern me about teacher predictions, is the accuracy of some marking throughout the year, if you know what I mean. So I know that some teachers are really harsh in their marking, some are really generous, some are examiners, some aren't. […] I think that's a massive concern. And unless you're an examiner I think it's really difficult to know how to mark accurately, if you know what I mean. So that would be something that I'd worry about in terms of the accuracy of the predicted grades. **head of department, further education establishment**

Some participants thought about scenarios in which they would be more trusting in the grades submitted by centres. These largely included the use of some kind of external standardisation of the judgements.

> If you organise it well, so if you have a system like Germany has in place […] as long as it's well standardised and staff are trained, I think that [teacher judgement] is a nice alternative. **head of department, further education establishment**

> In some ways this process is probably more fair than the exams, but we need to have something that is standardised across the country. **deputy head of centre, comprehensive**

This senior leader explained how it would have been useful to be able to align themselves with the standards set in other schools, at a national level.

> Looking backwards, retrospect, it's really easy to say [what could be improved] isn't it, so more time, greater opportunity to standardise, more shared conversation that was national. So, as I've explained, we were able to able to behave as a trust and take our shared understanding of subject standard, if that had been national that would have been amazing. So if there had been some way that that could have been managed that would have been brilliant. **senior leadership team member, academy**

Finally, interviewees reflected on the reliability of grades arrived at through exams in comparison to the centre judgements. Some commented on how they thought exams were more rigorous and objective. This was driven by an acknowledgement that exams are standardised, in that they are administered, marked and graded in a consistent manner.

> I think the toughest part though is that you need an external examination for fairness. I think that's the most key part though. Without that external part, then it just, it doesn't create a fair process. You need that objective person to assess a student's work and say OK, this is how they're going to do. **teacher/tutor, sixth form college**

> I think it's caused a lot of concern and I think it's so much easier […] when they take their exams. I used to think I would much prefer it if they just did coursework and this and that, but actually it's so much easier because it's objective. They know they sit an exam, they know there's nothing they can do about it. **teacher/tutor**, further education establishment

We further explore discussion on the validity of centre judgements versus exams in the following section.

6.2 Perceived validity of centre judgements

Validity of the centre judgements focuses on the degree to which the grades arrived at through this process reflect the students' true capabilities. It must be remembered that the judgements were intended to be a measure of how students would most likely have achieved in their final assessments, not a measure of how able they were in general. However, some of the discussions covered in this section do relate to the wider question of what is being measured.

6.2.1 Judgements versus exams

Aspects relating to the validity of the grades obtained through centre assessment in comparison to exams were considered. Some interviewees reflected how in the centre judgements the evidence used to contribute could be wide and varied, and considerate of each individual student's circumstances. They also liked that work at multiple time periods could be taken into account.

> I kind of like it [judgements], because I feel like in the case of that boy that I was telling you about where he struggles with interpreting the language of exam questions, I feel like I know how much geology he understands, and I know that it's more than that other girl whose language skills were better. And so I felt like I could rank him higher and feel good about that, because I know how much he understands about the subject that wouldn't necessarily come across in an exam. **teacher/tutor, sixth form college**

> [Centre judgement has] forced it into a much more meaningful exercise in terms of arriving at that grade through a process of looking at the learners' abilities holistically through the process. […] it's more of a 360 degree appraisal of the learners' ability rather than how they'd do on a day, or based on how they feel on a particular day, which I think is quite arbitrary in many cases. **head of department, further education establishment**

Many teachers commented that a disadvantage of exams was that they capture only a snapshot in time and are unable to measure capabilities over a more extended time period. This meant there were no contingencies for if a student had a 'bad day' on the day of the exam, or if the exam sampled questions that didn't play to their strengths.

> Even when a child goes into an exam one day, and they're fully prepared, on that day they might have had an argument with mum at home, or might have not had chance to have breakfast or something like that, so they still could do poor in the exam. So there's still an

> element there where the child could be disadvantaged even though they're taking the exam. **teacher/tutor, academy**

A couple of interviewees acknowledged that the centre judgement process may be positively biased towards those who work hard but struggle with exams and felt that this process was fairer than exams to those students.

> And if my bias is for a student who's worked hard, I don't think that's a bad outcome, because my personal feeling is that the system of exams in Britain teaches students that you don't have to work hard for two years and then you can cram it all in at the end and get the same result and that's not the way the real world works. […] So I've felt for a long time that the exam process undervalues some skills that are important to students. […] I recognise that I may have some bias towards those students who work harder, even if they aren't as clever. And the exam is the other way around, the exam is biased towards students who are clever, whether they've worked hard or not. And so I think that I prefer it this way. **teacher/tutor, sixth form college**

Some respondents did prefer exams though and felt that it was important for students to have been able to prove themselves in an exam, rather than having another person being responsible for their grades.

> I think it's unfair the students haven't had a chance to do it themselves because they do different things from what you expect. Some do better and some do worse than you would expect every year. **head of department, comprehensive**

> Students need to prove their grades, they need to prove what they've done and they need to prove their worth and that needs to be in evidence with them. Because you just can't have somebody going in and going 'oh, that person is going to be a distinction, that person is going to be that'. Well, with what evidence? A student needs to go in and prove. **teacher/tutor, academy**

6.2.2 Perceived validity of judgements for sustained effort vs last minute revisers

A big issue for many of the staff we spoke to was the issue of how to make allowance for the different profile of effort made by students. This has already been touched on a couple of times but making allowance for those increasing their efforts towards the end of their course or planning to do last-minute revision was a recurrent

theme. Sometimes these issues again touched on issues of the perception of sex differences, as we saw earlier in section 4.1.4.3.

> I think it's unfair to the kids who really just shine in exams last minute under pressure. I mean, I know a lot of kids that I teach, not sounding sexist, but a lot of boys perform better in that condition than girls do and they're really good in exams, but they do very little all year and I mean, these are the kind of kids that may have failed this year because they don't do much in class, but then pull it out of the bag in the exam. **teacher/tutor, further education establishment**

> I had a lad in my class who is, he's really bright, and he chose history before the options forms even came out at the end of year 8. He loves it […] and he would sometimes coast, and you'd say to him sometimes:
>
> 'How much revision did you do?'
>
> 'Oh, I didn't'.
>
> 'Oh right, so you got this grade, just think what it would be like if you did prepare, how much better it could be, because you did lose marks here and here and here.'
>
> […] So he was coasting, and when it came time to do his centre assessment grades, I just looked at the data that we did have, and I thought: right, well I don't know that he would have put his foot down, or pulled his finger out and done that little bit better. **teacher/tutor, comprehensive**

We have seen throughout these interviews that many centres placed a heavy emphasis on evidence, that judgements had to be supported by evidence. One consequence of this approach was that in some centres not enough allowance could always be given to students who left it to the end to work hard. The comment that follows was echoed in a reasonable number of our interviews.

> I think those kids who […] wouldn't attend well, they'd hand in no homework, and they just do it. They manage to get by in an exam situation. They probably haven't done as well from the [centre assessment] grade process, because they haven't handed in enough work that was evidence, because they'd have left everything to the last minute. And they probably are kids who are very bright kids, but […] they're probably very lazy some of them. So I think they probably haven't done as well this year, those people who leave it to the last minute, because like I said, this was about evidence. **deputy head of centre, further education establishment**

Even where there were signs that students were starting to increase their effort and show an improving trajectory, sometimes the centre's process with an emphasis on certain pieces of evidence such as mocks meant that this could not be reflected.

> The student who doesn't work hard in mocks because they don't take them seriously. [...] You feel bad, I feel bad because it's some I've got, and certainly those two, they were improving considerably. [They] suddenly decided after Christmas they were going to work hard for the exams. You knew […] they were going to get good grades. But that is not reflected in the school, so we haven't got any evidence of that standard. **teacher/tutor, academy**

Sometimes the guidance was perceived by teaching staff to force centres down this route.

> We made a very clear decision that anything submitted after the 20th of March was definitely not looked at. So the ones who were last-minute revisers of course were disadvantaged by that whole system, but those were the guidelines. **head of department, further education establishment**

This last-minute effort was of course not an issue in all subjects. For example, in art and design last-minute revision was not as feasible.

> In art, you can't really do last-minute revising. [...] What you've got in your sketch book is [it], and they now collect the sketch books in at the start of the exam instead of the end of the exam. So even more so they've got to keep up to date with what's set at the time, rather than thinking 'well I'll start the exam and I'll keep working on that for homework'. **head of department, comprehensive**

Interviewees in other subject areas such as English and maths below also reflected how last-minute hard work was not always likely to lead to improvement.

> Generally, if they've been working at a B grade for a year of the course, for an essay-based one if they write at that standard, they are not probably, however much work they do the couple of weeks before the exam, they probably aren't going to change their writing style and suddenly go up to an A grade. It does happen but it's rare. **head of department, sixth form college**

> It's never 'it clicks in May and suddenly they're there'. If it clicks in May it's too late for them normally, I don't really see them pulling it out of the bag in any spectacular fashion. **teacher/tutor, comprehensive**

However, in most interviews this kind of progress was acknowledged and there were various attempts described to make allowance for these types of students. Some interviewees suggested that professional intuition played a part in making the judgements for late revisers, and that this was often in combination with knowledge about how this type of student had performed in previous years.

> Taking into account those with these late trajectories and those with these late, 'no evidence' children I would call them, where you just think in my gut I know that that child will pull something out of the bag on the day, because he is exactly like that child from last year, the child the year before, the child the year before that, etc. etc. They will do it because they're that sort of child. **teacher/tutor, independent**

> The very late workers, those who leave it until the last minute. Because objectively there is no evidence that they would get good grades, you just know whether they would or wouldn't, because you just know what's happened in previous years. **teacher/tutor, comprehensive**

Discussions involving multiple teachers' perspectives was a common method of trying to overcome this issue.

> It's versus that student who [...] might pull it out the bag on the day of the exam, but how would you know? It's those students, and it's those outliers who are just really hard to separate. [...] There were two students, one who I felt should have been much higher than [another teacher] did and one who [another teacher] felt should have been much higher than I did. So it must come into it, that kind of personal judgement, of course it does. I guess the more teachers that are involved the easier that becomes, and you can lean on what somebody says who taught them the year before. **head of department, sixth form college**

6.2.3 Other issues affecting the perceived validity of centre judgements

We saw in section 3.2.2 that teaching staff sometimes had doubts about the validity of the mock exam results as a strong source of evidence for judgements, particularly around the lack of standardisation across classes and centres. Some more specific issues were mentioned such as students who had under-performed in their mocks as

they were focussing their efforts elsewhere, such as on university entrance exams or interviews.

> You've got some students who are going for a lot of medicine exams in, or basically medicine interviews before Christmas, and also Oxford and Cambridge interviews as well. And they were putting a lot of their efforts into preparing for those university interviews, and their mock grades may have dropped, well did drop slightly in December, and some dropped more than others. **head of department, sixth form college**

Teachers suggested that some judgemental allowance was made in these kind of situations.

Some teachers also reflected on how students mature differently, and that for some students, concepts are not grasped until much later on in the school year. Concerns for these groups of students were around how the evidence available to contribute towards judgements was not a wholly valid reflection of their capabilities, in that some evidence would be below the standard they were, or could be, capable of. For these students, the teachers felt that exams would be a more valid measure of students' abilities than the centre assessments.

> I think it [centre assessment] is better for the kids who work consistently throughout the year, but I think the kids, the weaker kids tend not to do that [for them] it clicks just before the exam and then they can do it in the exam. I think it's stacked against that kind of learner. **head of department, comprehensive**

6.2.4 Fairness of overall grade inflation to students next year and last year

The issue of grade inflation in summer 2020 was discussed in some interviews. Some interviewees felt that having some grade inflation was fair to students this year, because they would already be at a disadvantage due to the potential lack of confidence in their grades by employers or Universities.

> Do you know I don't really think it is unfair - I think, I know that it sounds like it's unfair, but actually these students that are going to go through the rest of their life knowing that the grades they've got might not be trusted. And I think that's quite a weight for them because say the student in particular who, say he comes out with two A*s for maths and further maths like he should, but he knows that everybody will always doubt that because he didn't actually sit the exam. And I think that's a bit rubbish for them. And equally if he walks around with A grades for the rest of his life, he will also think but I could have got

an A* and nobody knows that. So their grades are always going to be either something they think people don't trust or something that they think they could have done better if they'd been given a chance. So I think just let them have that little bonus of being more generous if necessary. **teacher/tutor, selective**

I mean they've had a horrendous time and I think it seems only fair that if it's going to go either way it should be definitely inflated rather than deflated I think, because they've missed out on loads.[…] I don't know in 10 years' time if people will look at these exam results and say 'oh it's 2020'. People will write on their job applications 'it was 2020' in brackets, you know, COVID-19 pandemic. […] I mean it's more for university entrants isn't it, this is the biggie, but I think universities are going to be understanding as well. **head of department, independent**

In contrast one teacher recognised that employers may be more likely to take other things into account outside of grades.

So what if you inflate those grades for a few kids? Well in the grand scheme of everybody in the country we're talking about a tiny pinch in the ocean and yeah inflate them, they deserve it, so what? And also if you're applying for a job I'd like to think that an employer would look and go I'm comparing somebody that got so many 9s in 2020 with someone that got so many 9s in 2021 or 2019 and you're going to go oh 2020 that was the year they didn't separate the grades and take it a bit more on how well they perform in interview and look at their experience which people quite frankly should be doing anyway. **teacher/tutor, comprehensive**

However, others felt conflicted that grade inflation in summer 2020 would not be fair to students in other years, who would be competing with these students for jobs or places on further or higher education courses.

I would say no to be honest [grade increases wouldn't be fair], because I think [...] it sounds harsh, but I'm thinking about the other kids of all the other years who've worked really hard and that's what I'm thinking. I'm thinking about fairness. **teacher/tutor, further education establishment**

So where I've read on Twitter where people say, 'well they should all just get the centre assessed grades. It doesn't matter that that would put the percentage getting each grade boundary much higher. It wouldn't matter.' And I think it would matter, because that isn't fair on

those that have gone before and those that will come after. **head of department, comprehensive**

6.3 Summary

There were mixed views about the degree of confidence those interviewed had in the centre assessments. In general, most were confident in their own CAGs and rank orders. Throughout the discussions, many recognised that there were challenges to reliability and validity that could make the centre assessments vulnerable to inaccuracies. These were largely related to pressures, biases, the degree to which fairness was maintained within a centre, and the types and weightings of evidence used. While most were clear that these issues had been overcome in their own centre, there were concerns about whether the same care had been taken across centres.

Some further reported limitations in their confidence that the judgements reflected what the students were capable of. There was a view that this type of assessment would advantage some students, and disadvantage others. The degree to which additional amendments were made to their judgements without their oversight was also a large concern for some.

7 Discussion

The 54 interviews we carried out with teaching staff for this project achieved two aims. One was to reflect on 2020, and understand how this year's extraordinary circumstances were dealt with – in other words how centre assessment grades (CAGs) and rank orders were determined by teaching staff. The second was to continue Ofqual's interest in how professional judgements are made in educational assessment contexts.

We have long been interested in how difficult assessment judgements are made when the evidence is multi-dimensional, such as in examination marking and coursework moderation decisions. The production of CAGs and rank orders was a highly complicated task. Even where mock examinations had been sat, or predicted grades existed, these formed only the starting point of the decision making, and a variety of other sources of evidence were required to be evaluated and merged into the final output. We summarise the findings under a series of broad areas.

7.1 A diversity of approaches were taken

Having considered the guidance published by Ofqual and the awarding organisations, centres took a wide variety of approaches to making their judgements, with hardly any two approaches quite the same in our interviews. This partly reflected the wide variety of roles, centre types, qualifications and subject areas of our interviewees. As a high-level classification, there were two main approaches. The majority of our interviews described a heavily designed process, coming from the senior centre leadership, and disseminated through departments. The aim of this process was to impose a degree of standardisation across what departments were doing.

Within this approach the standardisation was largely imposed through either the sharing of centrally-compiled data or through the sharing of detailed centre-devised guidance and plans. The shared data was organised in such a way as to lead the departments towards the calculation of grades in a similar way. The shared guidance detailed what departments were expected to do, but gave them a little more leeway over the choice and weighting of evidence to use in their decision-making. The latter was a more common approach than the former. Finally, some centres did not specify a detailed approach for departments to follow, but rather told them what outputs they expected to receive. Departments or individuals were left to determine their grades using their best professional judgement.

In most interviews, use of data was very important, with a lot of descriptions of data files such as spreadsheets containing all of the available data compiled for each student. This data represented a profile of each student's work over time, and captured teacher judgements in the form of marked or graded work. As suggested

above, sometimes this came from senior management or a dedicated data analysis team in the centre, or sometimes this was compiled within departments, sometimes by individual teacher/tutors. Arranging the data like this eased decision making, allowing comparisons to be made quickly and easily, both between students, and across time. Sometimes these data files contained some more qualitative information in addition to test and work marks to help with decision-making.

It was clear that this data-led approach came naturally to quite a lot of centres, due to the regular testing and strong data analysis they normally carry out on their students. They were already set up to make grade predictions, and to evaluate the accuracy of these predictions when results become available, using internal reporting and analysis, or using external data analysis tools and resources.

Although not a conclusive finding due to our small sample, it did appear as though the data-led approach was more likely to occur in larger centres, which may reflect the existence of dedicated data analysis resource within those centres. However, some larger centres, particularly in further education, took a more delegated approach due to the diversity of the qualifications they offered, making strong cross-department standardisation difficult.

Most of the interviews also described how the knowledge of individual students and their attributes were factored in. Often the data provided a starting point, but this set of grades and rank order would be modified based on student attributes. These were factors like whether students had made a lot of effort in key pieces of data/evidence such as mock examinations, individual circumstances leading to uneven performance, or their ongoing pattern of marks over time and how and when to extrapolate future performance from existing data.

In some interviews these more qualitative aspects were not mentioned, and a very heavily data-driven approach was taken, which was sometimes viewed by the participants as more objective, and therefore free from bias. However, the opposing view was that this would then be less fair as extenuating considerations and knowledge of the students as individual people were largely excluded.

When we consider the evidence used to make judgements, for most examination-based subjects mock examinations were the strongest source of evidence. In subjects with coursework or non-examined assessment (NEA) which included nearly all VTQs and some GQs such as art and design, this was one of the main sources of evidence to consider.

Although mocks were generally considered to be more controlled and standardised than other forms of centre-based assessment, and some centres had sat several mocks, various issues with the use of mocks were also discussed. Lack of standardisation across classes within a centre, variable marking, variable student effort, and lack of standardisation across centres were all issues making mocks

good, but not perfect, evidence. Other sources of evidence were always used to contextualise and refine the outcomes of mocks.

7.2 Teaching staff were extremely invested in trying to get this right

In all of our interviews, the teaching staff we spoke to were extremely invested in this, and spoke passionately about the need, and the pressure they felt, to get this right. This was almost universally because of the importance of their judgements on student futures, but they also reflected professional pride about their ability to do this well.

Quite a large chunk of many of the interviews was spent describing some of the specific circumstances they faced and difficult decisions they had to make for individual students, and how they had grappled with them and reached a decision. A lot of these ultimately were decided either by data in the form of previously marked work or assessments - the evidence had to be the final arbiter - or through some kind of qualitative judgement. These were also the source of many of the hard discussions that took place within departments when agreeing the merging of CAGs from several classes.

Teaching staff spoke widely about pressures. As well as the personal pressure to get this right, described above, they were also aware of external pressures, from parents and students wanting to know results or trying to influence judgements, to pressure from the media being critical of teachers. Mostly centres did a good job of shielding their staff from external pressures on their judgements, although the media, and social media was harder to shut out.

However, probably the major source of pressure on teaching staff at department-level was the perceived need to align their profile of grades with historic data. All our interviewees were aware that in GQs and some VTQs, a standardisation model was going to be applied to centre CAGs to ensure the year-on-year standard was maintained both nationally and within centres. Although the Ofqual guidance specified only that centres might want to use a comparison to previous years' data to check their CAGs, a lot of centres in our sample wanted to avoid being adjusted through standardisation and there were various approaches described to try to achieve this.

In most of the interviews we heard clear evidence that previous years' data, often including adjustments for the prior attainment of the cohorts involved, was being used. Many centres generated analysis of previous years' results centrally and shared this with departments at the start of the process, with various degrees of firmness from using these as actual quotas, to being for information to inform the judgements. Within departments, sometimes heads of departments, or individual

teacher/tutors would run this analysis to check how their CAGs compared to the results in recent years. Sometimes the analysis was done centrally by management, and then used to check initial CAGs from departments. Again, different degrees of firmness were applied in taking this data into consideration. In many cases, department-level staff felt a great pressure from management to lower some of their initial CAGs, and generally tried to resist, citing as much evidence as they could that their initial CAGs were valid.

This pressure to lower CAGs to meet prior performance data was usually unpopular, and generated stress for many individuals outside of senior management. Particularly unwelcome were cases when more senior staff imposed changes on the CAGs of individual teaching staff or those agreed at department level, often without consultation, and sometimes entirely unseen by those staff. More junior teaching staff sometimes only discovered the changes later on, sometimes post-submission. They felt that these lowered grades were unjustified or imposed on students by senior staff who did not know the students and what they could achieve. Of course, we must recognise that in general teacher judgements were generous, for the variety of reasons described in this report.

The interviewees rarely feared that there were any biases present in the submitted CAGs and rank orders. We did not always hear evidence of formal checking, for example by using data analysis, perhaps because this was done out of sight of class teachers/tutors, but mostly there seemed to be confidence that teacher expertise and fairness would not generate bias. Many did speak about their experiences of teaching a wide variety of types of students in a fair and open way. There were some centres using data analysis, and many shared information or some training to try to eliminate bias.

One final problem most interviewees discussed was the difficulty of making judgements across students with different effort profiles over time. The particular focus were those students who were showing signs of making additional effort either post-mock, just before lockdown, or were potentially 'last-minute revisers'. For all these individuals, the evidence and data collected could not really reflect an improving trajectory or their overall potential.

Some students had been showing improvement over a slightly longer period, so that the data captured signs of their improving results which could to a degree be projected forwards, but for other students there was no evidence whatsoever. These types of students caused great dilemmas. Some centres took a data-led approach which said if improvement wasn't in the data, or there was no verifiable evidence, then last-minute improvement could not be factored in at all. This was viewed as just unfortunate for those students who hadn't tried hard earlier in the course. Other centres were more flexible and allowed teaching staff to factor in some adjustment for late improvement, although usually these adjustments could not fully reflect

potential, as moving such a student above others with clearer evidence of achievement was not considered to be fair.

7.3 Submitted CAGs were of variable generosity across centres

There were different views on how well outcomes should match prior performance data. As described above, many centres informed their judgements with this data, or adjusted them centrally based on this data. There was also widespread awareness of the statistical standardisation model that would be applied for GQ and some VTQ. Some centres quite clearly wanted to match previous years' results perfectly, and so avoid any external standardisation.

Other centres were happy to allow a slight rise in outcomes, in the belief that there would be some kind of tolerance before statistical standardisation would be activated. Many of these centres had some uncertainty about how the statistical standardisation would operate, and whether it might lead to some unpredictable outcomes. This was one reason they wished to avoid being standardised.

A small number of centres were quite happy to see substantial increases. In some cases this was because they felt they had strong reasons to see an improvement from previous years, more often in specific departments, sometimes centre-wide. In others this was because the statistical standardisation was going to be applied anyway so it was out of their hands.

One area of tension in the judgemental process was the perceived need to match previous historical outcomes, versus wanting to provide grades which represented student performance on a good day. Historical results from examinations or final assessments always include students who under-perform on the day. These students have the potential to achieve a grade but fail to do so for a variety of reasons such as questions on papers they had not revised or exam stress. It was almost universally agreed in the interviews that under-performance is much more common than over-performance, although over-achieving can happen in some subjects, for example when a fortunate selection of questions match what a student has revised. In making judgements, this under-performance could not be predicted – it was by its nature unpredictable.

Working out which of a group of students with the potential to achieve a particular grade if all went well would be the ones to only achieve a lower grade was impossible – but it is exactly this that happens every year in assessments and is reflected in a centre's historical results. Therefore, a common and difficult situation was trying to decide which of a group of students with similar ability would be placed in a lower grade, or at the bottom of the rank order within a grade (in the knowledge

that statistical standardisation was likely to lower their grade). This was an unpopular and stressful task for our interviewees.

However, as mentioned above having grades reduced by management without consultation was even more unpopular. Ultimately, some students believed to have a moderate chance to achieve a particular grade were given a lower grade in the CAGs because centres wanted to match their historical results.

7.4 Different issues in general qualifications compared to vocational and technical qualifications

Most of our interviews were with teaching staff involved in GQs, and many of the themes picked up and described throughout this report are predominantly GQ issues. However, it is important to consider how the judgement process worked for those involved in the much more diverse VTQ landscape.

Firstly, it is worth drawing the distinction between the difficulties faced. Most GQ difficulties related to how to decide the correct grade and a fair rank order for students, because this was an exercise in prediction of how students would have achieved in the final assessments. For VTQ this was a much more evidence-based judgement, because students had large quantities of complete, or partially complete coursework, some of which had been marked or internally moderated. Even where exams were to be sat, some students had previously had an attempt at the exam. Therefore, there was very little concern about the strength of evidence available for placing students in a grade in VTQs. Rank ordering, where required, was still difficult, but most staff had actual work to hand to look at.

In VTQs the standardisation was often carried out through a quality assurance process, with awarding organisations sometimes sampling student work, or often just confirming the existence of sufficiently robust evidence on which centres had based their decisions. Statistical standardisation was only relevant for some VTQs. Therefore, the centre judgements had a different status to those in GQ, as they were perceived to be based on more evidence, and were thought to be less likely to be changed by awarding organisations, given that centre staff were used to assessing students internally, with external quality assurance, as their normal way of working.

Instead, in VTQs, the major difficulty was to do with limited timescales and lack of clarity. Differences in the requirements and the process across awarding organisations confused and added to the workload of VTQ teaching staff. The difficulties of actually submitting the judgements on some awarding organisation computer systems were also frequent complaints.

7.5 Anticipating statistical standardisation

Most of our interviews included discussion of the anticipated effects of statistical standardisation and the possible stress of results days. Teaching staff were worried about the lowering of student CAGs in the calculated grades and the impact this would have on individuals' futures. They also saw the pressure this could put on them from parents and students when the results were released. This was of course one of the considerations for centres when submitting their judgements. If the students and parents were to be told their CAGs, it was perhaps better to be seen to be a little generous, and have an external body have lowered them, than be seen to be harsh. The teaching staff could at least then face the students and parents and say they believed the student was worthy of a higher grade.

Most of our interviewees expected or wanted to see some 'leniency' or 'tolerance' in the application of standardisation, particularly where there were specific contexts that were not taken into account by the standardisation model. Many were also confident that their grades were accurate and would not change after standardisation.

7.6 Overall views

In broad terms, the teaching staff we interviewed were confident in their own judgements. They thought their approach resulted in valid and reliable judgements for their students, without much risk of bias. The only concern were cases where senior managers had, in their view, arbitrarily reduced some of the CAGs to match previous years' performance, and therefore given unjustifiably low CAGs.

However, many of the senior managers we spoke to felt that this was necessary, and sometimes that this was the right thing to do otherwise there would be no equality across centres, or fairness for previous or future years if CAGs were inflated. There was a certain amount of distrust in what other centres were doing. Occasionally there were worries that if other centres had not been as thorough as them, that could introduce biases and undermine the national centre judgement exercise. The reverse view was also in evidence, that this was an opportunity to show the nation that teacher judgements had integrity.

There was a mixed view from the interviewees of what the judgements were supposed to represent within GQ. A few felt that the centre judgements gave a more rounded, and therefore more valid view of students' abilities, and that this process will have been helpful to the kind of students who under-perform in exams. Of course, the intention of the judgements was to represent the most likely outcome for students in their final assessments – usually exams – and so the benefit to this kind of student would not necessarily be appropriate.

However, this distinction did not come out strongly in the interviews, and we have already discussed how attempting to predict who was going to under-perform in

exams was considered very difficult. There was concern around the under-grading of those students showing late improvements in performance for whatever reasons, and the slight threat this introduced to the validity of the judgements.

Overall, there was a view that using centre judgements was probably the best approach that could have been taken, given the extreme circumstances faced nationally.

7.7 Limitations

Our sample was designed to capture a wide variety of experiences, rather than represent the weight of experiences across the national teaching population. Therefore, we did not attempt to describe numbers or percentages of particular factors coming up in our interviews. This analysis gives an idea of the breadth of experiences and views, not their frequency. The companion survey (Holmes et al, 2021) gives a better insight into that kind of information.

The interview sample had a significant proportion of more senior members of staff, from senior leadership positions to heads of departments. These individuals would have had a greater involvement in the design of the process, and therefore may have been invested in the process employed within their centre. Therefore, we may have seen an emphasis on confidence in the process across many of our interviews. However, we have tried to also reflect the uncertainty of class teacher/tutors around adjustments made to their initial CAGs based on direct knowledge of the students, at department level, or by senior management, seen or unseen.

The interviewees would mostly have had only partial sight of the entire process within their centre. Senior staff would often have had less sight of the way individual judgements were made, while class teachers/tutors might have had only partial knowledge of some aspects of the process, including quality assurance by senior management. Whilst the interviewees mostly talked about what they directly experienced and knew had happened, in some cases there may have been some assumptions or guesswork concerning what occurred in the centre outside of their direct experience.

The interview sample may also represent a group of very engaged, motivated teaching staff, who were also confident of the validity of their judgements, and willing to talk to the qualification regulator about this; the sample did not represent so well newer, less experienced staff members. However, the overriding sense from the interviews was the care with which they had carried out this task. The passion with which they talked about individual cases and how hard they had tried to be fair was obvious. Even if they represent the more dedicated end of the spectrum this gives some confidence that most teachers took this task very seriously.

7.8 What can this tell us about centre/teacher assessment more generally

It is clear that centres took different approaches to the grades they were producing. Given the circumstances within which they were operating in GQ (and some VTQs), with the knowledge that some form of statistical adjustment was likely to take place, all approaches would be acceptable providing that the rank order they submitted had fundamental integrity. All of our interviewees felt this to be the case for their centre judgements.

Of course, circumstances changed, and when the decision was made to award the higher of the CAG or the calculated grade to each student, the variety of approaches did lead to some lack of comparability across centres. Unfortunately, the only possible standardising adjustment that could be applied to those grades was statistical in nature, which was judged to be unacceptable, so there was then no way to align the CAGs from different centres once they had been submitted.

The idea that mocks on their own could be used as a way to award qualification grades was problematic. For example, differences in marking, testing conditions, choice of papers and use of access arrangements would have made cross-centre comparisons very difficult. In some cases, even cross-class comparisons within a centre would have been difficult.

Looking ahead to the teacher assessed grades process in 2021, some of the pressures and stresses reflected on by teaching staff in this report will likely be present because of the importance of these grades for students. However, the situation is not the same. There has been disruption in schools and colleges for a considerable period of time, and while few might have assumed teacher assessed grades would be required prior to January 2021, many centres would have at least considered the possibility and contingency planning might have occurred. The time window available in 2021 is also a little longer, and a certain amount of time and effort will have been saved through the experience of 2020 and lessons learnt from that.

While the grading of students remains a high-stakes activity, the experience of 2020 will be useful to all involved in planning for 2021's teacher assessed grades. One issue this study cannot directly shed light on is the need for comparability of teacher assessed grades across centres. Ideally there will be a commonly understood concept of quality and scale to provide a sufficiently effective way to align judgements across centres through quality assurance, to check that they had been applying similar constructs and criteria in their judgements.

A final broader consideration in teacher assessment in general, is that there would need to be agreement about exactly what is being judged. Is it most appropriate or

valid to assess a students' ability to perform under normal assessment conditions? Or is it most valid to assess the student as a whole, evaluating their overall ability outside of their ability to perform in exams, but also factoring in their individual attributes? What is it that we would want to assess when deciding student futures?

7.9 Concluding reflections

We want to end this report with a few reflections from our interviews on the difficulties of this process in terms of the basic idea of judging performance in assessments that never took place. Although largely from a GQ perspective, these thoughts are relevant to the examined parts of VTQs, and also apply to the incomplete parts of internally-assessed coursework.

Perhaps the hardest part was that the whole process was never going to be able to produce grades that perfectly matched performance in exams due to the predictive nature of the task.

> [The] main problem is that you just […] don't really know what's going to happen in the future. You don't know […] what that student would have done between March and May, June when they did the exam. So many things, so many variables. **teacher/tutor, academy**

There is also the simple randomness around future exam performance which could never be fully factored in.

> But predicting grades is really quite difficult and […] teachers don't always get it right and there are statistics on that. And that's partly, I think, unpredictable marking, […] partly it's really difficult to predict how someone's going to do on the day. And I think yeah, I think that's what I've learned from this scenario. […] Teachers ranking people by ability, that's fine, but predicting [future exam] performance is actually [really hard]. **head of department, independent**

Despite the intrinsic impossibility of ever doing this task with 100% accuracy, what follows are two quotes reflecting on the difficult circumstances everyone faced.

> A global disaster happened and I think we've come up with the best circumstances that we could in the situation. How much fallout we're going to get, that's the thing, that's our abiding anxiety: we feel like we've done the best job we could, but we don't know how much fallout there's going to be from it. **deputy head of centre, independent**

I think it was the least worst option, like it's not ideal, but nothing's been ideal, so I don't see how you could improve it […] you weren't over-prescriptive about the data that you required or the evidence that was required, so I don't think you could have really improved it. […] It's the best it could have been considering the circumstances.
teacher/tutor, comprehensive

References

Allal, L. (2013). Teachers' professional judgement in assessment: A cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 20-34. https://doi.org/10.1080/0969594x.2012.736364

Begeny, J. C., Krouse, H. E., Brown, K. G. and Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psychology Review*, *40*(1), 23-38. https://doi.org/10.1080/02796015.2011.12087726

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation*, *17*(3), 141-159. https://doi.org/10.1080/13803611.2011.597112

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 69-90. https://doi.org/10.1080/0969594x.2012.703170

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education*, *7*(4), 279-301. https://doi.org/10.1207/s15324818ame0704_2

Cameron, T. A., Carroll, J. L., Taumoepeau, M. and Schaughency, E. (2019). How do New Zealand teachers assess children's oral language and literacy skills at school entry?. *New Zealand Journal of Educational Studies*, *54*(1), 69-97. https://doi.org/10.1007/s40841-019-00133-4

Cizek, G. J., Fitzgerald, S. M. and Rachor, R. A. (1995). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment*, *3*(2), 159-179. https://doi.org/10.1207/s15326977ea0302_3

Department for Education. (2020a). *Direction to Ofqual on GCSEs, AS and A levels*, 31 March 2020. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/

attachment_data/file/877611/Letter_from_Secretary_of_State_for_Education_to_Sally_Collier.pdf

Department for Education. (2020b). *Direction to Ofqual on vocational and technical qualifications*, 9 April 2020. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879024/Letter_from_Secretary_of_State_for_Education_to_Sally_Collier_-_VTQs_-__9_April_2020.pdf

Department for Education. (2021). *Letter from Gavin Williamson to Simon Lebus*, 13 January 2021. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/951797/Letter_from_Gavin_Williamson_to_Simon_Lebus.pdf

Everett, N. and Papageorgiou, J. (2011). *Investigating the accuracy of predicted A level grades as part of 2009 UCAS admission process*. London: BIS. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/32412/11-1043-investigating-accuracy-predicted-a-level-grades.pdf

Gabriele, A. J., Joram, E. and Park, K. H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction*, *45*, 49-60. https://doi.org/10.1016/j.learninstruc.2016.06.008

Gill, T. (2019). Methods used by teachers to predict final A Level grades for their students. *Research Matters: A Cambridge Assessment Publication, 28*, 33-42. Retrieved from https://www.cambridgeassessment.org.uk/Images/561974-methods-used-by-teachers-to-predict-final-a-level-grades-for-their-students.pdf

Gill, T. and Benton, T. (2015). The accuracy of forecast grades for OCR A levels in June 2014. *Statistics Report Series*, (90). Retrieved from https://www.cambridgeassessment.org.uk/Images/241261-the-accuracy-of-forecast-grades-for-ocr-a-levels-in-june-2014.pdf

Gill T. and Chang Y. (2013). *The accuracy of forecast grades for OCR A levels in June 2012. Statistics Report Series No.64.* Cambridge Assessment. Retrieved from https://www.cambridgeassessment.org.uk/Images/150215-the-accuracy-of-forecast-grades-for-ocr-a-levels-in-june-2012.pdf

Gill, T. and Rushton, N. (2011). *The accuracy of forecast grades for OCR A levels. Statistics Report Series No. 26*. Cambridge Assessment. Retrieved from https://www.cambridgeassessment.org.uk/Images/111066-the-accuracy-of-forecast-grades-for-ocr-a-levels-.pdf

Harlen, W. (2004). *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. Retrieved from http://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/ass_rv3.pdf?ver=2006-03-02-124720-170

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20(3),* 245-270. https://doi.org/10.1080/02671520500193744

Hoge, R. D. and Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297-313. https://doi.org/10.3102/00346543059003297

Holmes, S. D., Keys, E., Churchward, D. and Tonin, D. (2021). *Centre Assessment Grades: Teaching Staff Survey, Summer 2020.* Coventry, UK: Ofqual. Retrieved from https://www.gov.uk/government/publications/centre-judgement-research-survey-and-interviews-summer-2020

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education*, *28*(1), 91-105. https://doi.org/10.1080/02671522.2012.754229

Kahneman, D. (2013). *Thinking fast and slow.* Penguin, London

Lee, M.W., Walter, M. (2020). *Equality impact assessment: Literature review*. Coventry, UK: Ofqual. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879605/Equality_impact_assessment_literature_review_15_April_2020.pdf

Lindahl, E. (2007). *Comparing teachers' assessments and national test results: evidence from Sweden* (No. 2007: 24). Working Paper. https://core.ac.uk/download/pdf/7107168.pdf

Martínez, J. F., Stecher, B. and Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment*, *14*(2), 78-102. https://doi.org/10.1080/10627190903039429

Ministry of Education (2011). *Overall Teacher Judgement*. Retrieved from https://nzcurriculum.tki.org.nz/Archives/Assessment/National-Standards-archives/Fact-sheets/Overall-teacher-judgment

Murphy, R. and Wyness, G. (2020). Minority Report: the impact of predicted grades on university admissions of disadvantaged groups. *Education Economics*, 1-18. https://doi.org/10.1080/09645292.2020.1761945

Ofqual (2020). *Guidance for Heads of Centre, Heads of Department and teachers on objectivity in grading and ranking.* Coventry, UK: Ofqual. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/886921/Guidance_on_objectivity_in_grading_and_ranking_21MAY2020.pdf

Poskitt, J. and Mitchell, K. (2012). New Zealand teachers' overall teacher judgements (OTJs): Equivocal or unequivocal?. *Assessment Matters*, *4*, 53. https://doi.org/10.18296/am.0103

Rimfeld, K., Malanchini, M., Hannigan, L. J., Dale, P. S., Allen, R., Hart, S. A. and Plomin, R. (2019). Teacher assessments during compulsory education are as reliable, stable and heritable as standardized test scores. *Journal of Child Psychology and Psychiatry*, *60*(12), 1278-1288. https://doi.org/10.1111/jcpp.13070

Südkamp, A., Kaiser, J. and Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743. https://doi.org/10.1037/a0027627

UCAS. (2013). *Investigating the accuracy of predicted A level grades as part of the 2010 UCAS admission process.* UCAS. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/255060/bis-13-1088-investigating-the-accuracy-of-predicted-A-level-grades.pdf

UCAS. (2017). *End of cycle report 2017: Qualifications and competition*. Cheltenham, UK: UCAS. Retrieved from https://www.ucas.com/data-and-analysis/ucas-undergraduate-releases/ucas-undergraduate-analysis-reports/2017-end-cycle-report

Wyatt-Smith, C., Klenowski, V. and Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, policy & practice*, *17*(1), 59-75. https://doi.org/10.1080/09695940903565610

Wyness, G. (2016). *Predicted grades: accuracy and impact, a report for University and College Union.* London: UCU. Retrieved from https://www.ucu.org.uk/media/8409/Predicted-grades-accuracy-and-impact-Dec-16/pdf/Predicted_grades_report_Dec2016.pdf

Annex A - Glossary of terms

For clarity, we provide this glossary to clarify the meaning of some terms used repeatedly throughout this report.

AO

Awarding organisation. An organisation recognised by the qualifications regulators in England, Wales or Northern Ireland to develop, deliver and award qualifications. When delivering general qualifications, they are often also called exam boards.

CAGs

Centre Assessment Grades. This term includes any grades or rank orders required by awarding organisations for the awarding of qualifications. Some qualifications required grades only, some required grades and rank orders, while some only required rank orders. These submissions could be at whole-qualification level, or for individual units or components. We use the words 'grades' or 'rank orders' when talking specifically about those elements of the CAGs

Centre

Umbrella term for all types of institutions/organisations delivering the qualifications for which CAGs were required. Largely these are schools and colleges of various types, but also training providers and other educational establishments.

Cohort

The population of students in each academic year. The term may also be used to refer to a specific group of students – for example, a specific age group (the 16 to 18-year-old cohort) or to a group of students sitting a specific qualification (the GCSE maths cohort).

GQ

General Qualifications. GCSE, AS and A levels, and also IGCSE, Pre-U and EPQ. These qualifications adopted similar CAG submission requirements.

Grading

The process of grouping students into overall performance categories, usually on the basis of the marks received.

Grading: AS/A level

A levels award a grade, from highest to lowest, of A* (A level only), A, B, C, D and E, with a grade of U for unclassified.

Grading: GCSE

GCSEs award a grade, from highest to lowest, of 9, 8, 7, 6, 5, 4, 3, 2, 1, with a grade of U for unclassified.

Grading: vocational and technical qualifications

Vocational qualifications have a variety of different grading scales (for example, pass, merit and distinction).

HEI

Higher education institution. Often, but not always, a university.

Lockdown

March 16th 2020 when schools and business premises closed and unnecessary social contact ceased

QA

Quality assurance. The process within each centre where the CAGs determined by class teacher/tutors were checked for consistency, for example looking at student CAGs across subjects and compared to previous years' attainment data. A variety of different approaches were used and could be applied within departments, or within the centre senior leadership team.

SENCo

Special Educational Needs Co-ordinator. Staff with a specific remit to support students who require additional support within schools. In assessment terms this support includes extra time, scribes or assistive technology in examinations. SENCos are often class teachers.

SLT

senior leadership team. Members of centre staff who are responsible for the daily planning and management of the centre. These include the centre leaders and deputy leaders, such as the head teacher or principal, and assistant/deputy heads of centres and vice principals, key stage leaders, senior data staff and leaders or subject areas. SEND experts may also be part of SLT. Many SLT members can, but do not always, have a teaching role.

Statistical standardisation

The process of CAG adjustment applied by exam boards and awarding organisations under the direction of Ofqual to ensure the qualification-level outcomes for each centre were in line with the last 3 years of results for that centre. While in GQ the same process, devised in outline by Ofqual was to be used, in VTQ the quality assurance of CAGs was bespoke to awarding organisations and qualification types. For most VTQs there was no statistical standardisation, rather the CAGs were quality assured in a variety of ways by the awarding organisations, including checking the plausibility of a set of centre results by looking at previous results, or checking for the existence of sufficient evidence on which to base CAGs. Statistical

standardisation was put in place by a small number of VTQ awarding organisations using their own methods to produce 'calculated grades'.

Training Provider

An organisation or centre that provides training services for vocational and technical qualifications

VTQ

Vocational and Technical Qualifications. A very wide variety of qualifications requiring different CAG submissions. Different awarding organisations took different approaches – for example, some required qualification level CAGs and others component/unit level CAGs - and even within organisations there were differences between qualifications. Note that not all VTQs had CAGs – a large number either had adaptation applied to their assessments so that they could continue, or were delayed – put on hold until the lockdown ended.

Annex B – Interview schedule

Demographics – name, job title, centre type, subjects taught, which ones generated CAGs for, role in the process.

SENIOR LEADERS / HoD only

1   Can you tell me a little about your centre such as the variety of qualifications offered, the size of the cohorts taking qualifications, types of learners and the awarding organisations you had to engage with?

Prompts:

Largest entry qualifications

Any resitting cohort normally

Relative focus on GQ and VTQ

The Stages and preparation

2.   At a high level can you walk me through the different stages of the process you used to generate CAGs from start to finish?

Prompts:

Were there:

Prep/planning discussions/training

Generating class CAGs – any progress meetings

Qualification-level co-ordination/checking

3.   How was this process decided on? Was this a department- or centre-level decision?

IF THERE WERE PREP-MEETINGS inc during the grade generation period

4.   Can you tell me about the preparation before/during the time that individual student CAGs were worked on? In other words, what kinds of things were discussed, and what information was shared?

FOR SENIOR STAFF INC HoDs only

5.   Was the process top-down and implemented by departments, or was it designed by departments individually

FOR SENIOR STAFF INC HoDs only

6.   What consistency of process and information was there across departments?

FOR SENIOR STAFF INC HoDs only

7.   Did you submit CAGs for more than one awarding organisation? If so what differences were there between AOs?

Generating CAGs for individual students

8.   Thinking about individual students, how were individual students CAGs generated?

Prompts:

sources of evidence used?

most important evidence?

rank orders – was the process/evidence different?

HoC/SENIOR LEADERS only

9.   How much was the process used to generate individual students grades up to individual departments?

10.   How did you ensure the grades were fair such that a student better than another would have that reflected in their CAG/rank order?

11.   How did you ensure that this was based on their attainment/potential in exam and not on other things such as student personality characteristics?

12.   Do you think it is possible to disentangle student behaviours and characteristics such as engagement, attendance etc from potential attainment?

Bias issues

13. What (other things) did you and your centre do to ensure your judgements were as objective and bias-free as possible?

    Prompts:

    Guidance from Ofqual/awarding organisation/other sources

    Particular focus of discussion/training?

    e.g. Continuous sustained effort vs last minute revisers

    Protected characteristics – ethnicity/SEND

14. Did you have any specific discussions and/or training around the avoidance of bias?

    e.g. Guidance from Ofqual/awarding organisation/other sources

Other wider influences

STAFF WHO GENERATED CLASS CAGS ONLY

15. Did you feel any external pressures on your grade judgements?

    Prompt:

    Was anything done within your centre to minimise this?

SENIOR LEADERS only

16. What did you do to protect your staff from external influences on their judgements?

17. Were you aware of any social media coverage around this and did it have any impact?

18. Did you (or your staff) have any contact with students or parents who were anxious about this process? Did they have specific concerns?

Combining CAGs across classes

19. How were the class-level CAGs combined into one submission (if there was more than one class)?

    Prompt:

If you had to rank order, what difficulties did this introduce?

Did knowledge of the statistical standardisation affect your judgements, and in what way?

20. How did you ensure the grades were fair such that students of the same ability/attainment, even if in different classes would be given the same CAG?

Other sources of info:

comparison to other schools

comparison to previous years

what they think would happen in statistical moderation and was this incorporated into decisions

data from e.g. fisher family trust,

Submission

SENIOR LEADERS only

21. How closely involved in putting the qualification submissions together were you?

How did you ensure you had sufficient oversight of what departments were doing with individual submissions?

22. What is your confidence in the CAGs your centre has submitted?

23. How fair do you think these CAGs are for students of all types

Prompts:

e.g. Continuous sustained effort vs last minute revisers

Protected characteristics – ethnicity/SEND

Post-submission thoughts

24. What are you anticipating in regard to grade adjustments following statistical moderation or quality assuring by awarding organisations?

What are you hoping for, what's realistic, what's your worst case scenario?

What would be considered legitimate as a change?

Scenarios: CAGs & grade increases of 10%/20% at key grades – is this ok or do you think it would be unfair to the cohort next year, or even last year?

25.　What were the main problems you faced (if not already covered)?

26.　How could the CAG process have been improved?

Prompts

easier for you

fairer for students

27.　How did you feel about this whole process?

Emotional burden?

28.　Some press coverage has suggested this is a better process than exams. What is your view on this?

29　Anything else CAG-related you'd like to talk about