CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

What do we know about the evidence sources teachers used to determine Teacher Assessed Grades?

Research Report

Sylvia Vitello
Tony Leech
5 December 2022

Author contact details:

Sylvia Vitello & Tony Leech
Assessment Research and Development,
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

sylvia.vitello@cambridge.org
anthony.leech@cambridge.org
https://www.cambridge.org/

As a department of the university, Cambridge University Press & Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard.

We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team:
Research Division

If you need this document in a different format contact us telling us your name, email address and requirements and we will respond within 15 working days.

How to cite this publication:

Vitello, S. & Leech, T. (2022). *What do we know about the evidence sources teachers used to determine 2021 Teacher Assessed Grades?* Cambridge University Press & Assessment.

Contents

Executive Summary

- In summer 2021, as exams could not take place, GCSE, AS and A level grades in England were awarded by teachers, in accordance with relatively broad official guidance.
- This guidance stressed that grades had to be based on evidence of candidate work, though what this was, how much was needed or where/when it should come from were not tightly specified. This was to deal with variations in teaching and learning across centres as a consequence of the variable impact of the Covid-19 pandemic.
- The quality of these teacher assessed grades (TAGs) was assured by awarding organisations by sampling a selection of the evidence used.
- We looked at OCR samples for GCSE Mathematics and English Language, to try to get an understanding of what this evidence looked like at different centres, how it varied, and how different centres combined evidence to come up with final grades.
- The data we inspected was hugely varied in terms of the detail centres offered on what evidence was used to determine grades and how it was brought together. There was also considerable missing information. This constrained the analyses we could conduct and limited the conclusions we could draw.
- However, the data still provided us with valuable, useful insights into the TAG process.
- In particular, we found evidence of both frequent similarities between approaches, and also significant variations between and within centres in terms of what they did.
- Features of the assessment evidence we looked at, and the variations we found, are shown in Table E1.

Table E1. Features of the assessment evidence found within the submissions.

	Frequently observed (No features were universal)	**Variations observed (Between or within centres)**
Assessment evidence		
Type	• Exam-style assessments used • A combination of different exam-style assessments (e.g., full paper, half-paper, single exam question)	• Exclusive use of full exam papers • No use of full exam papers • Non-exam-style evidence (classwork, homework quizzes) used
Origin	• Use of OCR GCSE materials • Specific use of 2019 and 2020 GCSE materials, including Additional Assessment Materials	• OCR 2017, 2018 or legacy GCSE materials used • AQA or Pearson materials used • Materials not from awarding organisations (e.g., textbooks, Maths websites) used
Content coverage	• Candidates tested on broad subject content, covering all AOs covered in GCSE exams • GCSE Speaking AOs not assessed (English only)	• Exclusion of content not taught from assessments • Inclusion in assessments of content not taught • More assessment of certain content areas (e.g., writing in English)
Amount	• Multiple assessments used • Approximately same volume of assessment as in a normal GCSE session	• Much more assessment than in a GCSE session • Much less than in a GCSE session • Much more of certain content than in a GCSE session (e.g., writing in English)
Assessment conditions		
Centre-defined	• Assessments described as taken under “exam”, “formal” or “controlled” conditions	• Assessment conditions defined as “high” control • Assessment conditions defined as “medium” control

'level of control'		• Assessment conditions defined as “low” control
Specific conditions under which assessment taken	*[Not enough data]*	• Assessments taken in exam hall (Maths only) • Assessments taken in classroom • Assessments taken in other locations (e.g., at home) • Combination of assessment locations • Invigilated or supervised assessments • Open book assessments (English only) • Closed book assessments (English only)
Duration	*[Not enough data]*	• Timed assessments • Timings aligned with normal GCSE exam timings • GCSE timings extended by centres due to students’ lack of exam experience • Timings decided by centres • Untimed assessments
Date	• At least some assessments taken in April or May 2021	• All from April/May 2021 only • None from April/May 2021 • Assessments from 2018-2021 used • All taken from a small time period • Taken from a wide time period

- It was difficult to discern clear similarities across the whole dataset in terms of how teachers made judgements of the evidence. Table E2 shows the variations we observed.

Table E2. Variations in centres’ judgemental processes found within the submissions.

	Variations observed (between or within centres)
Assessment judgements	
Marking and/or grading of candidate performance on assessments	• Marks and grades provided • Only marks or only grades provided • Neither provided
Marking or grading procedures	• Double marking • Script anonymisation • Within-centre and cross-Trust moderation of marking procedures • Use of mark schemes • Use of grade descriptors (English only) • Marking annotations on some scripts • Some student feedback (English only)
Which grade boundaries used	• Those from session from which exam paper originated • 2019 grade boundaries (centres’ choice of standard) used for non-2019 papers • Centre-modified boundaries to account for reduced content • Centre-derived boundaries when they did not previously exist (e.g., for sub-sections of the paper) (English only)
Purpose of grade boundaries	• To determine final TAGs • For sense-checking
Final TAG judgements	
Prioritisation of evidence	• Based on assessment characteristics ○ Prioritising overall results on full papers

	○ Distinguishing between results in different content areas (i.e., writing vs reading) ○ Prioritising evidence taken under higher levels of control (e.g., exam conditions, timed, unseen papers) ○ Prioritising more recent evidence • Based on individual students' circumstances or performance ○ Reduced emphasis on evidence where students had mitigating circumstances or inadequate access arrangements ○ Exclusively basing TAGs on students' highest performance or set of best grades
Combining performance information	• Explicit mentions of "holistic" approach to determining TAGs (English only) • "Best fit" approach being taken to combine results of different assessments (Maths and English), or results in different content areas (English only) • Evaluating consistency of performance across different assessments (Maths and English), or different content areas (English only) • Diverse factors being taken into account during the TAG decision (e.g., grade descriptors, consistency of performance, assessment conditions, discrepancies between results and progression of learning) • No combination needed (e.g., TAGs based on results of one full paper or on one grade derived from a set of marks akin to qualification-level grade)
Internal quality assurance	• Head of Department-led moderation • Internal or external quality assurance meetings

- We conclude that, while the TAGs process provided assessment outcomes to candidates in what was a difficult situation and that these grades were on the whole accepted by stakeholders and wider society (at least compared to the situation in 2020), there are questions about comparability of standards between centres because of the level of variation we found.
- We are unable, due to the nature of the data, to make conclusions about the impact of centres' different approaches on the grades their candidates achieved, either individually or collectively, or to make conclusions about the extent of teacher bias.
- We end with four recommendations for improving possible future teacher assessment processes to enhance consistency, efficiency and comparability of standards.
- Recommendation 1: In any future situation in which grades are to be awarded by teachers, centres should be required to provide information in more consistent, more easily analysable formats.
- Recommendation 2: In future situations in which centres must gather evidence to support grading, guidance on which evidence to use must be as clear and as explicit as possible for centres.
- Recommendation 3: Should centres be required to provide TAGs in future, specific guidance should be given to support them to provide detailed explanations of exactly how they drew together the assessment evidence they used into a final grade.
- Recommendation 4: More generally, in future, robust exam-style evidence should be more habitually collected within the course of study. This would embed contingency within the assessment system for situations when terminal assessment is not possible.

Introduction

Background

In the autumn of 2020, it was generally assumed, or hoped, by the UK government that the disruption of the spring 2020 Covid-19 lockdown, which had resulted in the closure of schools for three months, the cancellation of GCSE, A level and other exams and the awarding of grades by an alternative process, would not be repeated in 2021. The intention was for exams to go ahead in summer 2021, although in a modified form as the pandemic was not over and some social-distancing restrictions were still in place. But on the 4th January, after a sharp rise in Covid-19 cases over the Christmas period, a new lockdown was instituted, and GCSE, AS and A level exams were again cancelled (Roberts & Danechi, 2021).

The Department for Education (DfE) and Ofqual instituted a public consultation on the approach to the 2021 process, to which there were nearly 2,500 institutional and 100,000 personal responses (DfE/Ofqual, 2021). Responses came from schools, local authorities, awarding bodies, examiners, parents and students, among others. A key principle from the start was that there be no “algorithm”, in order to prevent a repeat of the 2020 situation where the grading of GCSE and AS and A levels had to be changed at the last minute due to public disapproval of the use of an “algorithm” to determine students’ grades. Instead, the policy decision was made that grades would be determined by teacher judgement (referred to as Teacher Assessed Grades, or TAGs). Students would only be assessed on what they had been taught and, therefore, centres were given some latitude to determine the evidence on which student grades would be awarded. Allowed evidence would include mock exams, in-class tests, non-exam assessments, question sets provided by awarding organisations, and so on.

During spring 2021, the DfE, Ofqual, the Joint Council for Qualifications (JCQ) and awarding organisations (including OCR) collaborated to design the specific process for receiving, checking and quality-assuring the TAGs. This quality assurance process was intended to be deliberately very light-touch. This was due to the fact that no preparations for a more thorough contingency system were set in place in autumn 2020. Therefore, the extent to which particular requirements for what assessment evidence could look like was limited. So was the extent to which any moderation procedures could be put in place at short notice.

This report details an investigation into some of the evidence provided to OCR by centres for quality assurance during this process. Focusing on OCR’s two largest entry subjects in the quality assurance sample (GCSE English Language and GCSE Mathematics), we discuss the types of evidence used by centres to support their 2021 TAGs. We discuss factors including the types and volume of evidence used, what the assessments were based on, the conditions surrounding the assessment including the dates they were completed by students, as well as the approaches centres went through to combine these pieces of evidence to arrive at final TAGs.

Official guidance to centres on the 2021 TAG process

Any analysis, evaluation or interpretation of the assessment evidence and of the approach that centres used to determine their students' TAGs needs to be considered within the context of the official guidance that centres were given. One of the key guidance documents was published by JCQ in March 2021, after being reviewed by Ofqual and the DfE, entitled: "JCQ Guidance on the determination of grades for A/AS Levels and GCSEs for Summer 2021: Processes to be adopted by exam centres and support available from awarding organisations" (JCQ, 2021). We summarise guidance within this document that is particularly useful for contextualising centres' TAG processes.

There were only two mandatory pieces of information that all centres needed to provide to awarding organisations during the TAG process:

> (1) a Centre Policy, including a full and a summary version; and
> (2) a grade for each candidate for each qualification including for endorsement components (e.g., spoken language in GCSE English Language).

The Centre Policy was a document in which centres needed to outline the process they would take for determining grades and ensuring they were "appropriate, consistent and fair" (JCQ, 2021, p.6) across all subject departments. JCQ provided a pre-populated template of this policy document that could be adopted or adapted by centres. Centres were informed that there would be an external quality assurance process undertaken by awarding organisations, during which every centre's Centre Policy documents would be reviewed, and then a subset of centres would be sampled after the submission of TAGs to check that they had implemented their submitted policy. Centres were, therefore, encouraged to keep both records of evidence used for determining TAGs and their rationales for TAGs, to support internal and external quality assurance processes and to provide evidence should students wish to appeal their grades.

Another large portion of this guidance document was focused on advising centres on the types of assessment evidence they should use and how to combine the different sources of evidence to arrive at a final TAG. JCQ organised this set of guidance around a five-step process that they suggested centres could follow to help them make grading decisions:

- Step 1 - Consider what has been taught
- Step 2 - Collect the evidence
- Step 3 - Evaluate the quality of the evidence
- Step 4 - Establish whether the proposed range of evidence is appropriate for all students
- Step 5 - Assign a grade.

This five-step process signalled to teachers the suggested order that they should undertake the different processes. In particular, the fact that "collect the evidence" came considerably before "assign a grade" highlights the extent to which the phase of the process involving the production of assessment material was intended to be causally prior to the grading process (in other words that a grade should not have been determined and then evidence gathered to justify it). This has important implications for our understanding of processes undertaken

by the centres in the quality assurance sample in terms of our trying to reconstruct why they may have made particular decisions.

For each aspect of the TAG process, the guidance primarily focused on outlining the key factors and features of evidence centres should take into consideration as well as providing approaches that centres could take. There was notable variation in terms of how prescriptive different pieces of guidance were; some examples of variation are provided in Table 1. The mandatory pieces of guidance were more akin to broad principles than specific instructions.

These kinds of principles were provided for many aspects of the TAG process, including on:

- the content to be taught and assessed,
- the type and amount of assessment,
- consistency of approach across the cohort,
- how to prioritise sources of evidence,
- the nature of the TAG judgement,
- use of historical student and centre data,
- quality assurance, and
- retention of evidence of student work.

JCQ did not prescribe the specific way that these principles should be enacted by centres, although many pieces of guidance were either explicitly described as recommendatory or were implied to be as such by the fact that JCQ placed conditions around the use of certain alternatives. For example, JCQ specifically highlighted the types of "recommended evidence" that Ofqual outlined in their document, "Information for heads of centre, heads of department and teachers on the submission of teacher assessed grades: summer 2021." (Ofqual, 2021b). This list included student work in response to assessment materials from awarding organisations and centre-devised work that reflected the specification and marking processes used by awarding organisations. JCQ contrasted these types of evidence with statements about what evidence could be used in "limited circumstances, where other evidence is not available or possible to create." The latter, and various other pieces of guidance (see Table 1), seemed to be about giving centres as much flexibility of choice and approach as they could, whilst still aligning with the broad principles and aims of the TAG process they had outlined.

As well as containing different levels of prescription, the guidance also contained varying amounts of detail. One reason is that this document was intended to be used in combination with other official documents from JCQ, Ofqual and awarding organisations, which provided further information on particular aspects of the process. This wider set of documents exemplified the intended process in different ways, in order to help centres understand the guidance, apply it, and ultimately determine the appropriate TAGs for their students. One example of this was how the principle of "holistic judgement", which policy mandated had to form the basis of the TAG, was explained to centres. JCQ provided worked examples of how to reach a holistic decision for different combinations of evidence and circumstances around the evidence and student (JCQ, 2021). They used a scenario approach that specifically highlighted how centres could deal with six conditions: accounting for contextual factors in evidence; replacing evidence due to exceptional circumstances; marks available but no

work; partially completed non-exam assessment (NEA); minimal evidence available; and private candidate.

Overall, the prominence of broad principles and alternative options within the official guidance documents provided to centres, reflected the fact that the TAG process was designed to be flexible. Centres needed to understand that they were allowed to vary in their approach from each other (within certain parameters) and that they had to tailor decisions based on their students' circumstances and the availability of the evidence on their students' performance. This lack of prescription, however, can lead to uncertainty and a lack of clarity around what is expected. For example, Johnson and Coleman (2021) found that teachers considered the official guidance on TAGs to be unclear. They reported that teachers said that the final guidance arrived too late to be helpful and was difficult to use. As a consequence of the focus on maximum flexibility, these teachers had concerns about some centres obeying only the letter, rather than the spirit, of the guidance (e.g., by being overly generous with their grades on the basis that quality assurance was light-touch), and therefore potentially disadvantaging other centres or groups of students. There were, consequently, significant concerns about the relationship between the guidance and considerations of equity and fairness to students. Similar points were raised in media coverage of the process during Johnson and Coleman's data collection period; this may have had the effect of amplifying negative opinions, but the fact that the points were similar seems to reflect Johnson and Coleman's findings.

Table 1. Examples of JCQ guidance statements (direct quotes) showing variation in levels of prescription in the guidance.

Aspect of process	**How prescriptive is the guidance?**		
	Mandatory	**Recommendatory**	**Flexible**
Content to be taught and assessed	• The evidence used to make judgements must only include the appropriate assessment of content that has been taught. • Heads of Centre will be required to confirm that students have been taught sufficient content to form the basis for a grade.	• The aim is to include evidence that assesses the student's ability across a reasonable range of subject content reflecting, where possible, all assessment objectives, as set out in qualification specifications.	• There is no minimum requirement of content that students must have been taught. • It is not necessary for every aspect of the specification to be assessed to arrive at a grade.
Type and amount of assessment	• Reasonable adjustments for disabled students and access arrangements should have been in place when evidence was generated.	• Ofqual's guidance on recommended evidence…includes: …assessment materials provided by the awarding organisation…non-exam assessment…substantial class or homework…internal tests…mock exams…records of a student's capability and performance…records of each student's standard of	• Assessments used might be produced by awarding organisations, third parties or they might be teacher-devised tasks. • Consider what evidence there is of student performance, potentially collected over the course of study. • In some limited circumstances, where other evidence is not available or possible to create, an oral assessment may be an appropriate form of evidence.

		work over the course of study.	• If some evidence of students' work is not available, the marks can still be used in determining the final grade.
Consistency across the cohort	• Each student must only be graded on their performance based on the subject content they have been taught. • The rationale for any exceptions must be documented by the centre.	• Ideally, the evidence used will be consistent across the class or cohort.	
Prioritising sources of evidence	• Consideration should be given to the following: Coverage of assessment objectives; coverage of content; authenticity…; level of control…; [and] marking [procedures]. • Due consideration must be given to all the evidence collected for each student. • Where they [reasonable adjustments or access arrangements] were not [in place], centres should consider using other evidence or take it into account when coming to their judgement.	• While there is no one type of evidence that takes precedence, evidence gathered in conditions that enable confidence about the authenticity of the students' work will give more confidence in the overall holistic judgement. • More recent evidence is likely to be more representative of student performance.	
TAG judgement	• Grades should be based on a holistic, objective judgement of the evidence. • Decisions about potential must not factor in the student's grades. • The grading process this year…should account for the context in which each student's evidence has been produced. • Grade descriptors and grading exemplification must be used to make holistic judgements. • Professional experience and judgment will form a key part of this process.	• The grading process this year is not intended to be a formulaic calculation.	
Using historical student and centre data	• A grade derived based on a predicted trajectory or target grade is not permitted.	• Used appropriately, data on historical student and centre performance can help support the internal	

		quality assurance process for assigning grades. • Centres are advised to consider the profile of their results in previous years.	
Quality assurance	• Teachers' grading decisions will be subject to a school or college's overall quality assurance processes. • Centres' internal quality assurance process will ensure that standards are appropriate prior to sign-off by the Head of Centre.		
Retaining evidence of student work	• Centres should keep records of student evidence…so it can be found if a student wishes to appeal their grade.	• It is important that evidence... including copies of the student's work where available and any mark records, is retained safely by the centre.	

Review of previous research literature on TAGs

After the government's announcement in early 2021 that exams would not go ahead and, again, after the 2021 TAG results were released, various stakeholders and other bodies highlighted potential concerns about the TAG process. These related to a number of aspects, including generosity and inaccuracy of grading under systems of teacher assessment, and the specific potential for teachers' grades to be influenced by unconscious biases held against certain individuals or groups, as well as challenges for teacher workload and wellbeing. It is important to highlight that, while many concerns were identified, some evidence suggests that not all were entirely justified. It must also be noted that most of the evidence has come from indirect research (i.e., not based on data from the TAG process).

Accuracy and biases of teacher judgements

Many stakeholders were worried that the system of teacher assessment for 2021 grading would be unfit for purpose, because of concerns that it would result in grades being awarded that would not be an accurate representation of candidates' ability. A literature review conducted by Ofqual (Lee & Newton, 2021) on systematic divergences in results between teacher assessment and test-based assessments was central to arguments on the appropriateness of the 2021 system. Lee and Newton highlighted, on the one hand, that the relative agreement by rank order of students "between results from teacher and test-based assessments is of a comparable level to the relative agreement between teacher prediction and actual achievement." (Lee & Newton, 2021, p.24). On the other hand, they also highlighted strong evidence of a tendency for teachers to be more generous when predicting grades (e.g., in the context of university admissions) and that this tendency toward generous grading is also seen, although less commonly, when teachers assess current attainment (which is the target of TAGs).

One particular concern about the accuracy of teacher grades that was prominent in debates was whether there could be systematic biases in grading that would favour or disadvantage certain groups of students. Various researchers raised the possibility for teachers to be affected by biases that they may not be aware of (often referred to as 'unconscious biases') when determining their students' grades. Here, unconscious bias can be taken to mean the non-explicit associations that individuals hold between particular groups of people and particular characteristics, such as laziness or argumentativeness. This could potentially result in structural biases against particular groups of students in terms of their results being lower than those of other groups, when compared to the awarding of grades by awarding organisations where the particular candidates' characteristics are not known by markers. Lee and Newton (2021) investigated bias in TAG results and concluded that empirical results were mixed for some characteristics, including gender and ethnicity. However, for other characteristics such as socio-economic disadvantage and special educational needs (SEN), there was more consistent evidence of bias against students from more disadvantaged backgrounds or with SEN.

Many stakeholders had also raised concerns about the potential bias of the system that ended up being used for the 2020 process, where students' grades were based on centres' predictions of their students' performance had they sat the exams (known as Centre Assessed Grades, or CAGs). Ofqual's equalities analyses of the CAG results, however, concluded that the production of grades in the 2020 process was not "compromised by bias

in centres' judgements" (Ofqual, 2020, p.9). Although CAGs are different to TAGs, in that the former were predictions of candidates' performance not directly related to assessment evidence while TAGs are judgements of students' current attainment based on specific evidence, the controversies around the 2020 process meant issues of potential bias and unfairness were uppermost in the minds of stakeholders, teachers and students for 2021.

Teachers' approaches to determining TAGs

Since the start of the TAG process, two research studies have been published that have aimed to understand the process that teachers went through to determine TAGs. One study was conducted during the process itself (Johnson & Coleman, 2021) while the second (Holt-White & Cullinane, 2021) was conducted retrospectively, via Teacher Tapp, and reported in a research brief by the Sutton Trust. It was retrospective in that information from teachers was gathered at the end of June after they had submitted their grades.

1. Johnson and Coleman (2021)

Johnson and Coleman's (2021) research into the extraordinary and challenging experiences that teachers faced during the Covid-19 pandemic was broad in scope. Johnson and Coleman (2021) used surveys, interviews and diaries to gather rich information about the experiences of 15 teachers in England during the second and third terms of the 2020-21 academic year (January to May 2021). These were teachers of Year 11 and Year 13 students in subjects including English, science and geography. While the study was not focused on TAGs specifically, it covered the period where the TAG process was being designed, released to schools and then undertaken, and thus the teachers' views gave important insights into several aspects of the approach these teachers' centres took to TAGs.

First, the findings provided insights into the planning that the centres undertook: some centres had already made assessment plans before the official guidance on TAGs was released, whereas other centres delayed making decisions about what to do on assessment until the release of further guidance. Second, teachers gathered the assessment evidence through various methods. All teachers in their sample explained that they used mock exams to inform judgements, and, for some, this was the main basis of their judgement. Some centres had done mocks in autumn, others carried them out soon after schools re-opened on the 8th March 2021. Other types of assessment evidence such as non-exam assessment (NEA), coursework material, in-class assessments, unit tests and other work were also used by some teachers. Some teachers gathered additional evidence for students below their internal target grades following their initial evidence gathering and grading. Third, schools established procedures to avoid bias and potential accusations of bias, such as double marking and moderation. Some teachers saw a tension between their dual roles as teachers and assessors, seeing these roles as inherently in conflict; in other words that "as a teacher they were supposed to be looking at their students as a whole, and trying to identify their potential and encourage them, whilst as an assessor they were supposed to be making judgements on limited information." (Johnson & Coleman, 2021, p.43). However, others highlighted that previous examining experience supported their ability to judge their students dispassionately.

Johnson and Coleman's study also revealed the effects of Covid-19 on teachers' priorities with regard to balancing time for teaching and assessment taking, which may have affected the amount of evidence teachers had to base the students' TAGs on. The findings showed that the pandemic and lockdowns resulted in changes in content coverage, including avoiding practical skills teaching in science, a greater focus on wellbeing and more attention to key subject content. For many teachers, content delivery slowed initially, then was rushed when in-person teaching resumed. There was less discussion and fewer group tasks during the lockdown. Gauging learning and giving feedback were harder during lockdown, with less affluent students in particular disengaging from learning. Smaller groups and older learners fared better with remote teaching, though SEND students had more difficulty. In addition, in autumn 2020, some schools also reduced their volume of assessment specifically due to the importance of teaching content. After the TAGs process was revealed in spring 2021, teachers then consequently needed to gather evidence in different or more concentrated ways. This meant that their workload increased.

Uncertainties about whether teachers were collecting enough, or the right type, of assessment evidence, also had large impacts on teacher wellbeing and workload. These effects were not occurring purely because of the TAG requirements; workload and wellbeing were also being affected by the transition to in-person learning, blended learning and dealing with lost learning. Student wellbeing, particularly of those who suffered from mental health issues, was a particular challenge, and it was felt that the uncertainty of the TAGs process harmed the wellbeing of exam students.

2. Holt-White and Cullinane (2021)

In a research brief released after the collection of 2021 grades but before results were released, Holt-White and Cullinane (2021) reported findings of a poll of 3,221 teachers carried out through Teacher Tapp in June 2021 into "the materials being used to assess students this year and what teachers' views are on the new [TAG] process". This study was set in the context of the process's implications for A Levels and university access. The survey found that the teachers mostly used three or four assessments, though a sizeable proportion of A level teachers set more than six assessments. Most teachers indicated that they were trying to use the most objective evidence of student performance they had available. 96% of polled teachers reported that they used assessments carried out under exam conditions, with 80% of teachers using assessments based on past papers and 63% using mock exams. Teacher-written assessments, classwork, homework and so on were each used by considerably less than half of teachers. In this data these percentages are not mutually exclusive; for example, it is not possible to tell what percentage of mock exams, were not carried out under exam conditions, or how many teachers used exam conditions assessments alongside classwork. However, it is possible to say from this evidence that on the whole teachers preferred to use assessments that were as close as possible to exams. These findings about the types of evidence used seem to agree with those of Johnson and Coleman (2021), which showed a predominance of mock exams, but it is difficult to directly compare their results as both studies described and categorised the various types of assessments differently.

The study reported by Holt-White and Cullinane also showed some evidence of differences between schools of different types. Specifically, they reported that independent schools were

more likely than state schools to allow candidates prior access to questions, mark schemes or open book assessments. Schools with more affluent students and with higher Ofsted ratings were more likely to use mock exams and past papers.

The findings of this study revealed that, on the whole, teachers and students were only somewhat supportive of the 2021 system. The majority of teachers (58%) were confident in the system, and this was consistent across different levels of school deprivation, though 15% said that parents had pressured them to change grades, and this was more common at schools with more affluent intakes (see also Lada-Wilicki, 2021 for discussion of some of these issues). Young people themselves thought that the 2021 system was going to be fairer than 2020's grading process. Comparisons to normal years were mixed, with 42% of respondents in the Teacher Tapp sample suggesting the 2021 system would be fairer, and 45% less fair.

This study raised wider themes including the fairness of different university admissions processes; the potential for grade inflation and consequent issues of between-cohort fairness; the wellbeing of students and teachers; and learning loss. It therefore highlights the significance of the exam system in determining young people's futures and consequently the importance of getting assessment processes right.

Current research

The previous research on TAGs drew attention to some aspects of the process teachers underwent when determining TAGs for summer 2021, which highlighted various similarities but also variations in approaches across centres. Both studies' findings were based on accounts from teachers about the processes they undertook. The present study analysed the actual evidence that was submitted by centres as part of external quality assurance processes. We did this as a way to gain further insights into what happened in practice, especially across a wider range of centres.

At the start of the research, we did not know the extent to which the centres' submissions would be a useful source of data for our research aim, because of the high level of flexibility centres were permitted. Therefore, this study entailed taking an exploratory analysis of the data. Given that the data was submitted for quality assurance, we were confident it could give us insights into both the specifics of the evidence used as well as the nature of the TAG judgements centres made (that is, how teachers combined evidence from the different assessments to reach a TAG decision). Previous studies have reported on the first issue, the types of evidence used, but no published research, to date, has primarily investigated the judgemental aspect of the TAG process. As the nature of the TAG judgement was also part of the quality assurance process, we sought to analyse the data to gain understanding into this critical final aspect of the TAG process too. For both aspects (the evidence and the judgement), our aim was to understand variation between centres and between candidates within the same centre.

Methodology

Data

We were given access by OCR to the area of the Cambridge University Press & Assessment file system that contained the samples of evidence that centres provided to OCR for quality assurance of TAGs. Data from 148 centres was stored in this location, which included evidence for a variety of subjects and for both GCSE and A level. These centres represented a subset of all OCR centres; the sampling of centres was based on a process agreed across awarding organisations, in which both a set of risk factors, including previous cases of malpractice and unexpected rises in outcomes, and a random element, were used to select centres. Each sampled centre was required to submit TAG evidence for up to three OCR-selected qualifications if these were offered by the centre, which were (1) an AS or A level, (2) either GCSE English Language or GCSE Mathematics, and (3) one other GCSE. As was requested, each qualification sample needed to include the work of five candidates (or the maximum cohort if this was fewer than five candidates), who were selected by OCR from the total number of candidates entered for the particular qualification at that centre. These had to include a candidate receiving the highest grade achieved at that centre, a candidate receiving the lowest, and three others.

We looked at all the available samples for GCSE English Language and GCSE Mathematics (OCR J351 and J560, respectively). These two qualifications were selected because they had the highest numbers of samples.

Analysis

Our research aims were to investigate this data to determine what we could about the evidence sources used to support 2021 Teacher Assessed Grades, and to understand this data's structure and patterns of use. We wanted to discover what variations existed between and within centres both in terms of what evidence was used to determine grades, and how different types of evidence were combined in different ways by different centres. The research was exploratory and iterative, such that throughout the research process we continued to rethink our ideas and how the data we were looking at helped, or did not help, to answer the questions we were interested in.

To address these aims, each researcher looked at one of the qualifications, though methods were decided together, and findings were cross-checked to ensure a consistent approach was taken to inspecting and analysing the data. We each created spreadsheets setting out features of the evidence being investigated, treating each centre separately. These were developed from our looking both at the sampled evidence itself – scanned images of candidates' assessment scripts, featuring responses to past exam papers or other materials – and also at Assessment Records (where provided), which were documents in which the centres set out which evidence they had used and their rationales for doing so. The Centre Policy documents, having been assessed by OCR in a different part of the quality assurance process, were not part of the samples we looked at as they were not subject-specific.

For both GCSE subjects, the spreadsheets contained information on the centre, evidence types used, assessment conditions, and the marks achieved by candidates on those pieces

of evidence. Because the centres' documentation and submitted evidence varied widely in terms of content, structure and organisation, we had to take a manual approach to identifying and collating all the relevant information about the sources of evidence. No automated extraction of information was possible.

The features we chose to record were identified in an iterative process and were derived from the data, and, since the data for Maths and English Language were somewhat different, the features recorded looked slightly different too. For example, there tended to be more, and more fully detailed, Assessment Records available in English Language than were present for Maths. The Maths assessment scripts tended to have more information on them identifying their origin as well as details of the items, marks and duration, and tended to be mostly based on exam papers, while the English Language materials more frequently included answers to individual questions or writing assignments of various origin. This point will be highlighted more in the Findings section, but is mentioned here to indicate that, as a consequence of these data discrepancies, small differences developed between the approaches taken by the researchers when categorising the material. In the discussions which follow, themes arising from both subject samples will be discussed, but taking account of differences where necessary.

One important aspect of the analysis was grouping the different assessments into types. We first looked at other frameworks for how to do this, hoping we would be able to utilise a taxonomy of types previously used in a similar context. These included the list of the types of evidence listed in the official TAG guidance, and the typology used in OCR Review Records, which were spreadsheets used by OCR during the quality assurance process in which Centre Quality Assurers determined what kinds of evidence were available from each centre. However, these taxonomies proved difficult to use for various reasons. In particular, these taxonomies did not provide definitions of their assessment types, which, combined with the inconsistent information from centres on the nature of their assessments, meant that we could not map the assessments consistently and unambiguously onto them. We decided we could only draw meaningful distinctions between three types of assessments in the data (full exam paper, exam-style assessments and other materials), as these were the only types we could most consistently identify in the samples of evidence we analysed.

Finally, it should be noted that the findings that follow are generally described in qualitative ways. This is because of challenges in determining what "one piece" of evidence was for centres, and the fact there was little consistency in this (e.g., some centres described two half-papers as one assessment, and others as two). For this reason, we have generally avoided providing precise percentages as these are less meaningful than describing the data we looked at in relative terms.

Findings

The data for 26 centres was available for GCSE Mathematics and 13 centres for English Language. In this section we, first, describe the characteristics of the centres whose samples we investigated. Then, we present findings about the sources of evidence teachers used to determine TAGs by way of discussing themes that developed from our analysis of both GCSE Mathematics and English Language samples. It should be noted that some findings were relevant to different themes and, therefore, may be referred to in more than one sub-section; in each case, we contextualise the finding according to the specific theme being addressed.

Centres' characteristics

Here we briefly illustrate characteristics of the centres whose samples we investigated. We identified the centres' numbers from their submissions and used data about these centres to determine their characteristics. We focus on three characteristics here, which are centre type, location (in terms of Region of England) and centre gender profile. This is both because this information is more easily accessible (by contrast, information on a centre's ethnicity profile or percentage of candidates in receipt of free school meals is more difficult to access and link) and because some of the potential issues discussed in both the literature review section above, and in the wider discourse around 2021, concerned the potential for bias or differential outcomes across these factors.

As Table 2 shows, the sample of centres contained a variety of centre types, and there was some difference between Maths and English Language distributions. For example, most Maths centres were comprehensives of one form or another. For English Language there were slightly more independent schools and selective academies and far fewer comprehensive centres.

Table 2. Centre types of the sampled centres.

	Number of centres	
Centre Type	**Mathematics**	**English Language**
Academy (comprehensive)	12	2
Academy (selective)	0	3
Comprehensive	7	0
Independent School	2	4
Sixth Form College	1	0
FE College	1	2
Other	3	2
Total	**26**	**13**

Table 3 shows the locations of the centres, and that they came from across England, although, again, there were some subject differences. For Maths, most centres were based in the Midlands and in the north of England, while for English Language the largest number of centres were based in London and the South East.

Table 3. Location of the sampled centres by region.

	Number of centres	
Region	**Mathematics**	**English Language**
West Midlands	7	2
Yorkshire and the Humber	7	0
North West	6	1
East Midlands	2	2
London	2	4
North East	1	0
South East	1	3
East	0	1
Total	**26**	**13**

In addition, all but one of the Maths centres were mixed in terms of their students' gender composition, the other being an all-boys' school. For English Language, 9 of the 13 centres were gender mixed, three were all-boys' schools and one was an all-girls' school.

It is difficult from this relatively small number of centres to make many general claims about differences between TAG approaches by type of centre or other centre characteristics, and so this information is presented mainly for background. However, it is useful to consider characteristics of this sample of centres when interpreting findings.

What was the centre submission data like?

One major finding of our research, which is relevant to both subject samples, is that the data we inspected was hugely varied. This was a highly likely outcome of the TAG process because allowance for variation was built into the official guidance on TAGs. Ofqual allowed centres considerable latitude in the materials they could submit for awarding organisations' quality assurance procedures. In our sample, we found relevant information about the sources of evidence that were used by centres and how they were used to inform TAGs in a diverse range of documents. This meant that we often needed to collate different pieces of information from different places in order to obtain the most comprehensive understanding we could of the centres' TAG processes.

Information about which sources of evidence were used by centres for determining TAGs was found in three different places:

(1) within documentation containing details about the assessments and process for determining TAGs at sample or cohort level, often referred to as Assessment Records,
(2) within documentation containing details about the assessments and process for individual candidates, and
(3) within the submissions of candidates' work (e.g., exam scripts).

Most centres provided documentation containing information about either the overall assessment approach they took across their sample (or whole cohort) of candidates or the specific approach taken for the individual candidates sampled for quality assurance.

Sometimes both cohort and candidate-level descriptions were found. The provision of these kinds of documentation was slightly less common in Maths than in English Language. Most of the centres we reviewed had also uploaded images of their candidates' work. In English Language, 11 of the 13 centres submitted copies of assessment evidence for five candidates (the maximum sample). The other two centres submitted evidence for one or two candidates, although one of these centres provided a data table containing assessment results for seven candidates. We could not find information in their data explaining the discrepancy between the sample requirements (five candidates), the number of candidates for which we found evidence (two) and the number of candidates in this table (seven). In Maths, 25 of the 26 centres submitted evidence for all five candidates. The other was also sampled for five candidates, but submitted evidence for four, having no assessment evidence on the fifth candidate who consequently received a U grade from the centre.

Not all centres, however, provided all of these pieces of information. For example, in some cases, records of the assessment results (e.g., marks or grades) were provided, but copies of the actual assessments or assessment scripts were not. Moreover, in the data we did find, there was a large amount of variation with regard to the content, comprehensiveness, level of detail, structure and organisation. This variation was present among both the centres' documentation and candidates' assessment scripts, and there were also differences between subjects as well as between centres in the same subject. There were also cases where there was no evidence from a certain candidate for an assessment that most of their fellow students had undertaken; there could be legitimate reasons for this, such as candidate illness or other absence, or because the information was no longer available, leading to inter-centre variation, but we often could not find information about these discrepancies in the centres' submissions.

There were particular challenges with analysing the candidates' submitted work. For many centres, it was difficult to determine the specific assessments that were taken by the centres' sampled candidates. This was because few assessments had cover sheets that clearly related the assessment to the centres' documentation using the same terminology, and, in most cases, candidates' scripts were uploaded without information about what the questions were or where those questions came from. There were also inconsistencies in the organisation and formatting of the submitted assessment scripts. For example, for some centres, especially but not exclusively in English Language, some important details about their assessment scripts could not be determined, as their front covers, which would ordinarily detail such information as the origin of the assessment, its duration and number of marks, were often not present. In other cases, centres created new front covers for assessment scripts, with, for example, their own school information on them, or information that was used to communicate with candidates; these were similarly problematic for us in terms of the information they excluded.

The different pieces of information and data submitted by centres were also differentially useful for the two subjects we analysed. For GCSE English Language, it was the centres' documentation rather than the candidates' submitted work that provided the most useful information to start analysing the sources of evidence that centres considered for TAGs. There were three main reasons for this. First, assessment documentation was found for almost all centres, although in different forms and with varying levels of detail. Second, the information in these documents was structured in a way that provided an organising

framework to make sense of the candidates' submitted work. This proved particularly useful because, as noted above, many of the candidates' scripts were missing key information about the assessments (e.g., the exam question that was being answered and the exam paper it came from) and the uploaded pieces of work were, in several cases, not organised by assessment (e.g., answers to questions from the same exam paper did not always appear after each other). Third, although the information in the centres' documentation did not always reflect the final assessments that were used for every sampled candidate, they gave us insights into centres' intentions with regard to their approach for determining TAGs.

On the other hand, for Maths, the candidates' assessment scripts were often a better source of information. For nearly half of centres, documents that described the centre's overall approach for the subject were not present, and, even where present, many were incomplete. In some cases, there were summary documents for individual candidates or tables of marks and grades for candidates. The assessment scripts were more structured and typically had front covers that specified important information such as marks and duration. Moreover, in Maths, the candidates' responses were generally on the same page as the question they were responding to (whereas in English sometimes it was not clear to what question a particular assessment was a response, as assessment responses were less likely to be written on the same paper as the questions). Finally, there were sometimes issues whereby the Assessment Records did not appear to correctly record the information that was present in the sample. Thus, for Maths, the information relied upon most was the scanned copies of assessment scripts. These were available for almost all the candidates in the sample.

The substantial amount of missing data and the high level of inconsistency in the centres' documentation and in candidates' submitted evidence put some limitations on our analyses and affected the interpretations we could draw from the data. In particular, we could not infer anything from a lack of evidence of a particular aspect of the process since the centre may simply have not recorded it. These data inconsistencies also made it difficult to produce quantitative information showing how similar the centres were in their use of evidence and approach to determining TAGs. This meant we had to take a qualitative approach to analysing the data, focusing on the specific details in the evidence for each centre in order to appropriately convey the commonalities and differences between the centres in the sample we analysed.

A consequence of the fact that centres varied enormously in the amount of evidence they provided was that the explicability of their grading decisions from that evidence varied too. This will be discussed in more detail later. While many centres wrote clearly in their subject Assessment Records about the rationale for why particular evidence was used and how it supported the decisions, there was no information for many others, and in some cases such information was partial. Many other features of the evidence discussed were not present or could not be determined for many centres' samples. This included the duration and number of marks on papers, the origin of the assessment material and the conditions under which papers were sat. While it is understandable that evidence samples would be differently structured, given the latitude that centres were given in terms of what evidence was permitted to support their grades, what is perhaps less defensible is that for many samples it was not clear how the awarded grades had been derived from the material, how different pieces of evidence were drawn together to determine such a grade, or whether another person would have come to the same grade based on looking at that material. Since we

wished to investigate these issues in our centres, the frequent absence of such information had an impact on our findings, which are consequently more tentative in this area.

What assessment evidence did we find?

Types of evidence

We classified the evidence that we identified into three types:

- full exam papers
- exam-style assessments, and
- other materials.

These three categories captured the range of evidence we found and reflected, to some extent, how centres themselves distinguished between different pieces of evidence. The nature of the submission data meant that our classification of evidence into these categories was often based on collating different pieces of information about those assessments and taking a best-fit approach.

As stated in the Method section, these categorisations were defined by us; we could not use pre-existing frameworks of types, such as the list of types in the official guidance, as these were not clearly defined and types of assessment within them could not be unambiguously distinguished from one another. In addition, it is important to highlight here that these categories may not have corresponded exactly to the ways that centres thought of their materials because we could not use centres' own labels for their assessments as a basis for classification. This is for a number of reasons. Firstly, centres did not always categorise materials in the same way; for example, one centre may have defined a paper as a "mock" and another defined the exact same paper as an end-of-unit assessment. Secondly, in some cases, centres' descriptions of (or labels for) their assessments were chronologically contingent. A "mock" exam that was sat in November 2020 may have been used ultimately as a source of TAG evidence and so could not be usefully described as a mock at that time (during the TAG generation process) as it was used for summative purposes. Thus, had we used centres' definitions of their assessment evidence, the categorisations would not have been meaningful across centres. Distinctions between the three types we decided to use (full exam paper, exam-style assessments and other materials) were the only ones we could most consistently identify for the samples of evidence we analysed.

In our classification, "full exam paper" refers to a student response to a full-length examination paper. For the OCR J560 GCSE Mathematics qualification, there are three exam papers: each one is 100 marks, is sat in a 90-minute period, and is not divided into separate sections that test specific content areas or topics. For OCR J351 GCSE English Language, there are two exam papers: each one is 80 marks, is sat in two hours, and features two sections both worth 40 marks, Section A testing Reading and Section B testing Writing. (There is also a Spoken Language Endorsement non-exam assessment.) The full-length papers in the sample of evidence we reviewed were either previous OCR papers from earlier exam sessions, alternative papers available to centres via a secure setting (those produced in case the main papers were compromised), or, in one case, a full paper from another awarding body.

“Exam-style assessments” refers to student responses to examination questions organised in ways other than the form of a full examination paper. For example, we found full papers with questions removed as particular content had not been covered during teaching, as well as papers put together by centres from questions originally written for a variety of different sessions’ exams, and half-length papers. This category also includes ‘Additional Assessment Materials’; these materials, which were produced by OCR and other awarding organisations to be used as assessments in 2021, used previous exam sessions’ questions and were set up as half-length papers.

Finally, a small number of other materials were submitted, including homework quizzes, worksheets, textbook exercises, and classwork such as worked problems. These were few in number, and very diverse in form. These types of materials were found within submissions of only a small number of centres. For three English Language centres, this evidence was only used for one or two candidates; some of the candidates had non-typical education circumstances (for example, they had an Education and Health Care Plan), while for other candidates this type of evidence was submitted as an alternative for assessments that they missed. The same was generally true in Maths, where most of the other material was used for a single candidate, or in other cases was described by centres as low priority supporting material. The apparent secondary function of other materials was supported by comments made by centres about the reason and role of these materials in determining TAGs. For example, one English Language centre mentioned a variety of non-exam materials (e.g., textbook fictional analysis activities) and referred to them collectively as “additional evidence”.

Overall, the vast majority of evidence in both subjects was based on pre-existing exam questions in one form or another, perhaps highlighting the confidence centres felt in the exam format as a method of assessment in this subject, or their lack of other materials with which to fairly assess their students. For English Language, some centres explicitly stated that they chose such assessments because they covered all the required Assessment Objectives (AOs) and marking could be standardised. Similarly, in Maths, the fact that such materials could be more easily standardised and involved all (or most) candidates sitting them at the same time, was highlighted. Moreover, one centre used such assessments to give candidates the “opportunity to complete an assessment in full exam conditions with access arrangements in place” and described this as “the closest form of assessment” to that of a normal year.

Across the whole sample for both Maths and English Language, it was clear that exam-style assessments were used the most frequently by centres (more than full exam papers), such that they are worth discussing in more detail. Around two-thirds of the pieces of evidence reviewed for English Language, and half for Maths, were of this type. For English Language, these assessments had 1 to 4 questions, and were worth between 10 and 56 marks. Most assessments contained a single question, which ranged from being worth 12 marks to 40 marks. A few of these assessments contained a specific subset of questions from OCR exam papers; for example, one centre chose to include questions 1, 3, 5 and 6 from an OCR past paper as an assessment. In some cases, the centres explained that this was so that the particular questions aligned with the content that the students were taught or because the questions gave students more opportunity to demonstrate their skills. In Maths, the exam-

style assessments were worth between 21 and 88 marks. However, the vast majority were around 45-55 marks – in other words, around half a typical exam paper. Some of these half papers were Additional Assessment Materials, which were designed by OCR specifically to be of that length.

There were additional notable differences between English Language and Maths in terms of how exam-style assessments were used, which suggested there were distinctions between how teachers in the two subjects thought about the constructs that were being assessed in their subjects. The Maths exam-style assessments were often just parts of full exams, which some centres constructed as such for practical reasons. For example, some centres said explicitly that they divided papers into two halves so they could be sat in two lessons' worth of class time. On the other hand, in English Language, where some assessments were based on a single question, there was some evidence that they were used to target specific skills. In particular, exam-style assessments more often focused on testing writing rather than reading skills. The finding that English Language exam-style assessments included more single-question assessments of writing could be interpreted as meaning that teachers needed (or wanted) more evidence of that skill to base their TAG decision on. Another possibility is that they had more of this evidence available, perhaps because this skill is assessed more often during the normal teaching process for this subject.

There was sometimes uncertainty around whether, in practice, evidence that looked like exam-style assessments were taken as full exam papers or vice versa, without being recorded as such in Assessment Records. For example, one English centre provided assessment evidence for two full papers only, but they recorded the evidence as two separate Section A assessments and two separate Section B assessments within their Assessment Record rather than as two full papers. We could not find sufficient information to confirm whether these assessments were taken as full papers in one sitting or as separate sections at different times but there was some evidence suggesting this possibility. The centre explained that each assessment was sat in the same month ("completed in May 2021") and taken under the same conditions: "Unseen paper. Taken in exam conditions in teachers' respective classrooms".

Centres' combinations of evidence types

While the previous section described which types of evidence were used across the whole sample of centres we reviewed, we now discuss the combinations of types used by individual centres. In our sample, some centres chose to use only full exams, while others used only exam-style assessments. However, it was clear that most centres in both subjects used both of these assessment types. Occasionally, as discussed above, other material was also used, though for specific reasons.

In Maths, four centres used only exam-style assessments, and two used only full exams. Of those using only exam-style assessments, one centre, for example, used a session's worth of half-papers, to allow candidates to sit assessments in classroom time and across a period of weeks. Meanwhile, those that used full exams highlighted the extent to which they wanted to replicate the experience of a normal exam session. However, the vast majority of centres used both of these types of evidence (in various combinations and to different degrees). For example, one centre used a full exam session's worth of three papers, sat in November 2020, alongside up to six half-papers per tier of exam-style assessments, and two additional

exam-style assessments, sat in April and May of 2021. This centre's reasons for using both kinds of assessments were practical, and to do with fitting assessment into classroom teaching schedules, which is discussed more in the later section "What conditions were assessment taken under?" Similarly, further findings on the amount of evidence that centres gathered is described in more detail in the section "How much evidence was there on each candidate's performance?"

Similarly to Maths, it was also rare for English Language centres to use only full exam papers to determine their students' TAGs. Only one centre in our sample explicitly stated that they exclusively used full papers for TAGs; in this case, the students took two full papers. For almost all the other centres, it was clear from their submitted evidence that they used a combination of full exam papers (at least one) and assessments comprising either one or more exam questions. When the English centres provided explanations for their choice of assessments, this often related to the coverage of the subject content within or across the set of assessments. Unlike Maths, a few English centres gave reasons that seemed to specifically distinguish between their uses of full exam papers versus exam-style assessments. One English centre clearly stated that full exam papers were deliberately used as 'benchmarks' for making judgements about students' overall TAGs. Other reasons that seemed related to the use of full exam paper was where centres explained that these assessments covered all the AOs, although they could have achieved this using multiple exam-style assessments and, therefore, it was likely that there were other factors to this decision.

Origins of evidence

A further area of interest was understanding the origins of the evidence used by centres – that is, in the case of past papers, which years they came from, and, in the case of other materials, their wider origin. For Maths, the origin of the evidence could be determined for roughly two-thirds of the material, and for English Language slightly more.

For Maths, papers from OCR's November 2020 session and both 2019 sessions were most frequently used, while more infrequently there was evidence of OCR's 2017, 2018 and legacy qualification (i.e., those prior to the reformed 9-1 GCSEs) papers being used. Additional Assessment Materials, specifically provided by OCR for the TAGs process, were also used quite commonly. It can, therefore, be seen that most of the assessments used came from the most recent exam sessions. Of 2019 material, only a very small fraction came from the November session in that year; the vast majority was from the June session. For English, the distribution was similar to Maths, with most of the material coming from OCR past papers and specifically 2019 and 2020 papers, though for a greater number of pieces of evidence the precise paper year was unclear. Additional Assessment Materials and older materials were also used by some centres. It was more common in English for the materials to feature questions from a mixture of different papers than for the Maths assessments.

A small number of pieces of evidence in the Maths and English samples were not based on OCR material. This included one Maths centre whose evidence was entirely based on responses to AQA Additional Assessment Materials, and others where classwork materials or quizzes were used from websites such as Mathsgenie.co.uk, Dr Front Maths and Kangaroo Maths. In English, the pieces of evidence that were not based on past exam

papers included a writing test planning sheet; a freewriting task; a collection of activities that contained workbook exercises; and a TAG and TAG justification for one student from teaching staff at a different academy to the one determining the final TAG. Some other pieces of evidence were mentioned in some English centres' documentation but were not actually submitted. These included assessments set by a former tutor, textbook activities, activities set in class each week to be independently completed and submitted, and classwork being used to replace missed assessments.

On the whole, though, it was clear that the majority of sampled centres preferred their GCSE assessments to be based on previous, recent, OCR exam material. This is likely to be the case because doing so effectively outsources (in an entirely appropriate manner) the construction of such assessments – in terms of issues such as specification coverage and level of difficulty – to the awarding organisation. Teachers may not have felt they had the skills or knowledge, or indeed the time and resources, to construct materials of their own, especially at short notice, and it is well established that in normal years these are awarding organisation roles.

Coverage of Assessment Objectives and subject content

The assessments used by centres covered different parts of the GCSE specifications. For Maths it was generally the case that most materials covered all of the GCSE's AOs. This is because all three papers in OCR's J560 qualification cover all three AOs – "Use and apply standard techniques" (AO1), "Reason, interpret and communicate mathematically" (AO2) and "Solve problems within mathematical and other contexts" (AO3) – in the same proportions.[1] It is not necessarily the case that assessments used for 2021 TAGs will have covered the AOs in exactly the same proportions as normal exams, due to some centres having removed particular content from assessments, but, since most assessments were either full or half-papers, it can be expected that the proportions were similar in most cases.

Coverage of AOs across Maths centres was therefore similar. However, in terms of the exclusion of specific topics from assessments, there were some notable differences between centres. Evidence from the documentation of some centres suggested this was because the teachers had not been able to cover certain topics in teaching the course, due to lessons being missed because of the pandemic. A number of Maths centres chose to remove particular questions, those covering topics they had not covered, from the assessments they gave to candidates. For example, one centre omitted four questions from an otherwise full exam paper, and graded candidates on their answers to the others. In other words, these centres focused on aligning their assessments to the content they had taught.

There were also other approaches taken by Maths centres to dealing with the relationship between assessment and content coverage in teaching. One Maths centre gave the whole of a former exam paper to their candidates but, rather than removing this content from the assessment, they modified the grade boundaries. Another Maths centre indicated in their Assessment Record where each assessment covered each of the twelve units of subject content (in addition to each AO). In terms of whether calculators were or were not allowed

[1] NB These proportions are different between foundation and higher tiers. Foundation tier has AO1 at 50% and AO2 and AO3 at 25%. Higher tier has 40% for AO1 and 30% for AO2 and AO3.

(where typically two of the papers have calculators permitted and one does not), centres tended to either give candidates the opportunity to sit papers roughly in this proportion, with more calculator papers, or, in cases where only two papers were sat, one of each.

Moreover, many Maths centres provided substantial, long lists of all the content they had covered during teaching, which was sometimes at the level of what particular candidates had been taught. This information was not linked to the assessments they used but may have been useful for centres' own internal quality assurance process, or it may have been submitted specifically for OCR's quality assurance process so that the centre could justify that they had taught sufficient content to award their candidate TAGs. For example, one centre recorded (for each candidate) which units of content that they considered that candidate had been taught satisfactorily and which not: in one case, Algebraic Proficiency and Mathematical Movement had not been taught. Another centre indicated in its Assessment Record that certain groups had covered the majority of the course and others less content; however, it did not define these groups in ways that had meaning to people external to the centre nor did this content coverage issue noticeably affect the assessments they chose to offer, except in one case where a question was omitted from a paper.

The situation in English Language was more varied. Across the whole sample, most of the assessment evidence covered reading and writing. Only a small number of pieces targeted speaking - one in each of two centres, both using assessments based on the GCSE Spoken Language Endorsement. Every centre (for which details were found) used assessments that, together, covered all six AOs from OCR's two exam components (J351/01 and J351/02). Both components assess reading and writing constructs but component 01 focuses on non-fiction texts and writing while component 02 focuses on literary texts and creative writing. However, variation among the centres with regard to both the balance of reading and writing and the balance of non-fiction and literary questions included in the assessments they used to determine TAGs was evident.

Only two English centres assessed the same balance of reading and writing AOs as would be covered in a normal session, assessing the equivalent of two of OCR's exam papers. All the other centres contained an unequal coverage of reading and writing AOs. Almost all centres included more assessments on writing than on reading. Indeed, half of the centres included extra writing assessments (in addition to full sets of exam questions). The other centres included extra writing and reading questions (in addition to the full sets of exam questions) but the reading questions were almost always a subset of the full reading section (Section A) of the OCR exam papers. Often, these centres included an extra Question 3 and/or Question 4 which are the two longest questions from Section A and assess candidates' text analysis.

In addition, almost every English centre also included questions from both English exam components, which meant they assessed understanding of non-fiction as well as literary texts. However, there was large variation in how many questions were drawn from each of the papers and, thus, how many times the content from each component was assessed. For example, one centre made a specific decision to assess content only from component 01 as this aligned with the content that students learnt during the year due to interrupted teaching. This was similar to the situation in a number of Maths centres, in which assessments were modified to align with the taught content. Another English centre also decided to feature

component 01 constructs more heavily in their assessments but for a different reason; they viewed this component as targeting concepts that were distinct from the concepts in the English Literature course, which they were also teaching. In effect, the centre had modified their view of the subject construct, seemingly deviating from the specification.

Other English centres made deliberate decisions to assess content from both components because of the distinction between non-fiction and creative writing between the components, to ensure that they assessed both types of writing. However, there was no evidence of centres drawing analogous distinctions between the reading skills of these two types of text. As another example, a different centre deliberately mixed questions from the two components in the same assessment to give students optionality of questions. Finally, in two other centres, coverage of the components varied between their candidates such that some candidates completed more assessments on component 01 and fewer from component 02, or vice versa. There was, therefore, evidence of centres focusing on specific components, ensuring they focused on both types of writing (but not reading) and having different approaches for different candidates within their centre.

On the whole, there was evidence that English Language teachers varied more in their understanding of the subject's constructs, and how they were expressed in the assessments they were setting their candidates, than was found among the Maths teachers. This is likely a consequence of the fact that all three Maths exams cover all the subject content and AOs, and as such it is easy for centres to simply set papers and be confident that they will cover what is required. The important feature, it seems, for Maths teachers was focusing on aligning their assessments to what they taught their students. For English Language, this was also the case, but the teachers were also ensuring that they had enough evidence of the different constructs that make up English Language (i.e., reading and writing, or assessment on non-fiction and literary texts) to assess their candidates against constructs from the whole specification. Another factor is that skills in English may take different amounts of time to develop or be more complex to evidence – thus justifying some centres' decisions to have, for instance, more writing assessments in their selections of evidence. Moreover, that there was more variation in English Language centres' approaches is an important finding in terms of the issue of the comparability of 2021 grades between centres.

How much evidence, in total, was there on each candidate's performance?

This next section discusses how much evidence centres used, in total, to inform their students' TAGs. We found considerable variation in what centres considered to be "one piece" of evidence in their Assessment Records (or other documentation), for instance in terms of whether half-papers were seen to be one or two pieces, and the extent to which collections of 'other material' were viewed as one item. Because of this, it turned out not to be appropriate, or meaningful, to consider the amount of evidence in this way, and, therefore, instead we addressed this question in relation to the overall volume of assessment. We specifically compared the amount of evidence used to evaluate candidates' performance for TAGs to the amount that would be used in a normal exam session. As previously stated, in Maths, an OCR candidate in a normal year would sit three 90-minute exams, each of 100 marks, while in English Language they would take two 2-hour exams, each 80 marks, and also take the Spoken Language Endorsement component. If candidate evidence provided for quality assurance for 2021 was less than this, questions might be raised as to the extent to which such evidence could provide adequate evidence of their

ability. However, it is also worth bearing in mind that, particularly in English Language, another element of the comparability to a normal exam session, in addition to marks and assessment duration, is the balance of content being assessed (i.e., both reading and writing, in equal volume). We have already discussed this aspect, to some extent, in the section on subject overage, where we noted that, for many English Language centres in this sample, they did not use the same balance of content in their evidence as in a normal exam session.

It was not possible, due to missing, varied and incomplete information, to have analysed the evidence for every candidate in the sample to determine whether they each had more or less evidence than would be presented in a typical exam session. As a result, we focus on case studies of centres to provide more comprehensive illustrations of the variation we found both between centres and between candidates within the same centre.

In Maths centres, it was typical for there to be, in general, the same volume of assessment as in a normal exam session, or slightly more. For example, at one centre, each of the four foundation tier candidates had exactly the same evidence in their samples – responses to six half-length papers based on dividing all three of the 2020 November OCR papers in half. As a result, each candidate's evidence was the same volume as would be the case in a typical session – 300 marks and 270 minutes' worth, and the same coverage of content and AOs as normal. There were also centres where one or more full sets of exam papers were used, and others where the cumulative total of various exam style assessments and full papers exceeded the amount of evidence created for a normal session. That said, there were a couple of centres where the amount of evidence was more limited, for instance to just two full papers per candidate, or four or five half-papers.

However, perhaps a greater source of variation was in relation to consistency between candidates in the same centre. For example, at one Maths centre, evidence was based primarily on sets of exam-style assessments created from OCR questions. However, the amount of evidence for each candidate in this centre's sample was different, with between four and eight half-papers for each. For two of the candidates there appeared to be more assessment overall (in terms of marks available) than in a typical exam session, and for the other three a little less. In another Maths centre the amount of evidence also varied sharply by candidate. One foundation tier candidate's evidence was equivalent to two full exam sessions' worth of papers (specifically this was sets of answers to both the June 2019 and November 2020 sets of papers). Due to absences, two other candidates provided only some of the evidence from these full sets: four and five out of the six papers respectively. Another candidate was only present for one of these papers and as such had only a single piece of evidence to support their grade (which was at grade 1).

Similarly, another Maths centre used an exam session's worth of full papers, for which their sampled candidates had between two and five results (some candidates sat both the higher and foundation tier sets of papers), and a set of smaller exam-style questions, for which there were between two and four pieces per candidate. Finally, a further Maths centre used a variety of evidence to support candidate grades, including full OCR exam papers from both J560 and legacy qualifications, as well as centre-created sets of exam-style assessments and classwork. One candidate's evidence clearly exceeded the volume of examination that would have taken place in a normal session, though the centre explained that this was due

to the fact that they sat on the grade 4/5 borderline (as reported by the centre) and therefore both higher and foundation evidence was presented. For three other candidates, slightly less material was presented than in a normal session. It can therefore be seen that variation in the amount of evidence presented for Maths candidates existed within centres as well as between them. In some cases, this variation was a consequence of candidates' varying availability to sit the assessments, but other reasons were also mentioned by centres.

For English Language, it was meaningful to assess equivalence based on the types of question and skills being assessed rather than simply the marks. This is because of the distinct structure of OCR's English Language exam papers, which contain a specific number and type of questions on testing reading and another distinct set on testing writing. A few centres used exactly the same volume of assessment as a typical exam session (i.e., two full papers). More commonly (just over half the sample), centres used assessments that amounted to the equivalent of two full papers plus some additional questions; in almost all cases these additional questions were between one and three writing questions. This volume was found even for centres that did not explicitly mention using any full exam papers: when considered together, the exam-style assessments were equivalent to at least one full exam paper, with most amounting to two full papers plus some questions. Two centres asked their students to take almost the equivalent of three full papers, although both fell short because one or two of the reading questions from Section A were not tested. A few centres gathered less evidence than typical of an exam session, but all tested at least one full paper plus additional questions. Again, in most cases they fell short because they were missing reading questions. In other words, whereas different pieces of evidence from the same Maths centre tended to cover the same content, in English Language it was more common for certain content types to be assessed more than others (e.g., writing more than reading) with a consequent impact on the overall volume of assessment.

As we found for Maths, for several English centres we also saw between-candidate differences in the total amount of evidence used. The availability and structure of the centres' documentation in English showed that for most cases the variation between the sampled candidates was due to specific circumstances surrounding individual candidates that meant they had to deviate from the general approach they set out. However, in one English centre the level of variation between candidates was much larger and, due to the lack of documentation, we could not identify whether or not there was an intended common approach across the cohort and what that was. The sampled candidates from this centre had assessment results from between one to four full exam papers plus two to six single writing questions, and the number of these writing questions seemed unrelated to how many full papers they took. There was no reason found in their documentation for why such different amounts of assessment evidence was used across their sample of candidates.

It was certainly the case that for many candidates in both Maths and English Language an abundance of evidence was presented, sometimes more assessment material than they would have been examined on in a normal exam session. It may be that some centres wanted to provide more evidence than they thought was strictly necessary, in order to be able to justify their grades to OCR. On the other hand, for a significant minority of candidates, evidence was sparse, raising questions about the extent to which it could be relied upon as true evidence of their ability. In these cases, it is reasonable to consider that the volume of evidence resulted from "whatever they had available". It is also possible,

however, that evidence of poorer performance was omitted, but we could not determine the extent to which this was the case. Moreover, there was variation in evidence quantity between candidates in the same centre too, raising further comparability questions. This all highlights the significant relationship between evidence availability, the awarding of grades and centre accountability for their decisions (and therefore the incentive to provide as much evidence as possible).

What conditions were assessments taken under?

Some centres provided information about the conditions under which their assessments were taken, including the assessments' level of control, duration and date taken.

Level of control

For just under half of all the assessments in both subjects there was some information about the level of control under which the assessments were taken. Most of these assessments were classified, by centres, using a three-part classification: high, medium or low control. In English Language, most assessments were classified as high control, with a smaller number in the other categories. The full exam papers were almost always described as high control, but exam-style assessments had more varied levels of control. In Maths, almost all of the assessments (irrespective of type) for which a level of control was given were classified as high control.

Even though the classification of assessments as being taken under high, medium and low control was used across centres, there was evidence, from centres' descriptions of the conditions for individual assessments, that the same definition of control was not used by all these centres. Some assessments with the same label had different characteristics to each other, and there was also some overlap between assessments labelled with different levels of control. Table 4 shows examples of terms used by English Language centres to describe their 'high', 'medium' and 'low' control assessments; this shows the similarities, differences and level of consistency between and within the categories. In Maths, there was less evidence of centres describing their conditions at all, and also less variation in levels of control, with most assessments being described as of a high level.

Table 4. How the levels of control were described by English Language centres.

Aspect	High control	Medium control	Low control
Assessment 'label'	exam; assessment; summer assessment; mock; presentation & discussion; test	Classwork	classwork; homework; scaffolded homework; exam
Exam conditions	exam conditions; formal exam conditions; full exam conditions; strictly controlled circumstances; controlled conditions	-	-
Where	exam hall; classroom; school	classroom; remote	remote; home
Supervision	invigilated; supervised by non-specialist teacher	supervised by teacher; invigilated	-
Preparation / resources	unseen paper; closed book; no pre-planning time or discussion	preparation beforehand; no preparation beforehand	preparation beforehand; scaffolding; preparation and discussion about assessment guidelines

Regardless of the way levels of control were classified, from Table 4 we see clear variation in the conditions under which assessments were taken. Assessments varied with regard to:

- whether “exam conditions” were followed,
- how the assessments were described (e.g., as exams, tests, homework, classwork),
- where the assessments were taken,
- who supervised the candidates,
- whether the assessments were taken under time conditions,
- whether the paper had been seen before,
- whether the assessment was open or closed book,
- whether preparation or planning was completed beforehand, and
- the level of teacher input (e.g., guidance or scaffolding).

There was also some evidence of assessments being used in a more formative way, especially for English. Some of the candidate scripts included feedback comments from the teacher to the student. We could not find any direct information about the purpose of this feedback, and whether it affected the teachers’ grading of the assessments or was used during judgements around TAGs. Nonetheless, the fact that it was present suggests that these assessments may have been undertaken under lower levels of control – perhaps relating to assessments being open book or involving pre-planning – and consequently that some of the assessments had more of a formative purpose. This would not be especially surprising, particularly for assessments sat earlier in the course, as at that time centres and students would not necessarily have expected that exams would have been cancelled. There may be questions for validity, however, in using assessments originally designed to be more formative in nature for summative purposes.

There was a small amount of evidence in both subjects of candidates being given repeated opportunities to sit assessments (or questions within assessments). We could not identify whether this was part of a deliberate approach on the part of the particular centres, or just a feature of what evidence was available.

Several centres in both Maths and English Language said that they attempted to replicate exam conditions as closely as possible; they described assessments as being undertaken under “formal” or “full” exam conditions or under “strictly controlled circumstances”. Some gave specific details of conditions that reflected those of exams such as assessments being closed book, using unseen papers, conducted in invigilated exam halls and so on. In Maths, while details of conditions were not given as frequently, the fact that most assessments were full exam papers or exam-style assessments undertaken under a high level of control implies that they were used to approximate the experience of a normal exam session, with alternatives to that used mainly on the basis of Covid-19 or other practical exigencies.

Again, this relates to the sense that centres were only temporarily taking on roles generally undertaken by awarding organisations – in this case, determining the rules under which exams should take place – and that they saw the ways that things typically took place as generally appropriate. However, the fact that some centres did things differently, and the fact

that they saw the levels of control differently, raises questions about the comparability of the evidence produced under these conditions.

Duration of assessment

There was sometimes information provided relating to how long the students were allocated to complete the assessments. In Maths, since the assessments were mostly full or half-papers, it was likely that full papers took the same amount of time as they would have done in a normal exam session (90 minutes), and exam-style assessments that were half-papers, half that. A number of Maths centres' Assessment Records confirmed that this was the case, though for other centres there was no direct evidence of this. A handful of assessments were of different lengths, for instance 30 or 60 minutes, but these were particular to individual centres. For example, one such centre created its own assessments to be 60 minutes' long, in order to cover only the content that their teachers had been able to teach most securely, while another centre used legacy qualification assessments, which were of a shorter length.

For English Language there was little information on assessment duration in the centres' documentation. As for Maths, several of the English Language assessments were described as being taken under "exam conditions", and therefore it is likely that the same timings were taken as for exam sessions. For OCR papers this would be 2 hours for a full paper, with one hour recommended for Section A (four reading questions) and one hour recommended for Section B (one writing question). For a few full papers and some exam-style assessments, the candidates' scripts showed evidence which supports this reasoning. Some candidates' full exam scripts contained front covers created by the centre, which confirmed the hours their students were given to sit the assessment – in all cases this was 2 hours. This was also the case for some exam style assessments comprising a single section of a paper, which contained centre-modified front covers that specified the timings; these aligned for the typical exam timings for each section (one hour for a Section A or a Section B).

However, there was some evidence of different durations than those used in live exam sessions. One English centre provided specific details about how they had modified the timings from OCR's timings and explained why. This centre set the reading section to 90 minutes and the writing section to 60 minutes in order to accommodate the fact that students had not experienced a full exam session before. Other English centres seemed to remove timings altogether. For example, one centre described one full exam paper as being submitted to a "strict deadline", which may suggest there was no limit on the duration as long it was completed by a particular date (although it is possible that a duration of how long to spend on the assessment was given to students but not recorded). Furthermore, we also found evidence of assessment durations varying between different groups of students in the same centre, such that the same assessment was done in timed conditions by some students and as homework (perhaps untimed) by other students.

In summary, then, while it appears that many centres set assessments to be the length of typical exam papers (or occasionally half of that), there were also interesting examples of variations to this. Centres seemed to take the opportunity of the freedoms provided to them by the TAG policy guidance to break up papers in order to allow them to be sat across multiple classroom sessions, or at different times, or to alter the amount of time available, presumably so that they could best reflect their candidates' abilities and utilise the resources

and time they had available. While this was likely to be of benefit to candidates, given these factors, there are inevitable questions to be asked about the comparability of evidence between centres.

Date of assessment

Finally, we inspected the evidence for information about when the assessments were sat by candidates. Across both subjects, we could identify the date of assessment for around two-thirds of assessments, and often these were only approximate (to a month). This means that for a large minority of the evidence we could not determine when it was gathered. For the assessment dates we found, we analysed them from two perspectives: (i) in relation to the release of policy and guidance on TAGs, and (ii) in relation to the GCSE academic schedule. The former enabled us to understand the extent to which students and teachers understood what these assessments were going to be used for when the assessments were taken, while the latter helped us understand the extent of teaching and learning at the time of the taking of the assessments. Throughout this section we refer to these two perspectives where relevant or informative.

Within both the Maths and English Language samples, we found that the largest number of assessments were taken in April or May 2021. This period was after the TAG guidance was published (in March 2021) and covers the last months of Year 11, and hence it is likely that the majority of teaching will have taken place by this point. This finding seemed to be the only overall similarity between the subjects with regard to assessment dates. In the Maths sample, another large proportion of assessments were taken in 2020, especially in the autumn term, while the least evidence came from the first three months of 2021 (January-March), which was a period of lockdown. It was interesting, if not surprising, to find more assessments from autumn 2020 than early 2021, given that the (high-level) TAG policy was only announced in January 2021. Hence in autumn 2020 it was generally assumed that exams would go ahead in 2021 (in some form). Thus, it is likely that the assessment evidence collected in 2020 reflected centres' typical process, that of using assessments in formative ways during this part of Year 11. That said, some centres may have anticipated exam cancellations given the increasing Covid-19 case rates at this time. Across the English centres the opposite pattern was found to Maths; few assessments were from the autumn term of 2020 and a large number were from early 2021, although it was not clear why this was the case. In addition, in English, the assessments came from a wider time frame than for Maths in that evidence came from the two academic years of GCSE study; several English assessments had been taken in Year 10 (i.e., September 2019 and August 2020). One English candidate even had assessment scripts from Year 9 in their evidence, but this was the exception.

The above discusses these patterns across the whole sample. However, it was important to look more specifically at what occurred at centre level. This was because centres used multiple assessments to inform their TAGs, which may have been distributed differently across the years at different centres. For example, students could have taken different assessments at different time points, or conversely the centres could have scheduled all assessments to take place at a similar point in the year. We found large variation between individual centres as well as, sometimes, between individual candidates within the same centre. We present descriptions of the dates of assessment found for different centres to

illustrate some similarities and differences, including commonalities between some English and Maths centres.

Across both subjects, almost every centre (for which we found assessment dates) used at least one assessment that was taken in April or May 2021 (the months closest to the TAG decision), although not all candidates in some centres had been able to take these assessments. The ways these assessments were scheduled also varied between centres. For example, one Maths centre set weekly assessments in April and May 2021, responses to all of which were used for TAG decisions. There was one centre (in the English sample) where none of the assessments were taken from 2021; in this centre the latest assessment was from December 2020 (i.e., the first term in Year 11). Apart from April/May 2021, there were no other time periods that were common across all centres, although there were some commonalities between centres within the same subject sample.

Within the Maths sample, many centres used only material from April and May 2021, and in some cases where earlier material was used this was only there for candidates who had a paucity of more recent evidence. For English, there were also centres that exclusively used assessment evidence from the 2021 spring or summer months. For one of these centres, every piece of evidence came from May 2021. However, this reliance on evidence from those most recent months was only the case for a small number of English centres. In our sample of English centres, it was more common to find the use of evidence from a wider time period. Some of these centres used evidence from earlier in Year 11 too, specifically the end of the autumn term (December 2020). More centres, however, used evidence from across Year 10 and Year 11. For example, one centre used two pieces of evidence from Year 10 (December 2019 and February 2020) and three from Year 11 taken between February and April 2021.

A few Maths centres used evidence beyond May 2021 too. The breadth of the time period also varied between these centres, as was found for the English centres. For example, at one Maths centre, assessments were chosen from two time points, May 2021 and November 2020, evidence from the latter of which may have been viewed as mocks at the time. Another Maths centre's evidence came from a much larger number of time periods spanning May 2020 to May 2021: two were from 2021 (March and May) and two from 2020, one of which was in Year 11 (November 2020) and one in Year 10 (May).

Furthermore, we found some evidence of variation around these dates for candidates within individual centres, with some students in the centre having sat assessments earlier or later than either stated in the documentation or than found for the other students. In one English Language centre, variation was specifically built-in to the classroom structures, with teachers being given flexibility to administer assessments when they judged it to be appropriate.

Overall, it was clear that obtaining recent evidence of students' ability in the subject was important for centres, although how they supplemented this with other pieces of evidence from other time points varied widely across the centres. This aligns with the official guidance which highlighted, as an important priority, that teaching should continue "for as long as possible" – and that, therefore, assessment should be pushed as late as possible. The guidance also highlighted that "more recent evidence is likely to be more representative of student performance," though with exceptions including when students suffered ill health

during later periods. In interpreting these findings, it should be borne in mind that for much of the assessment evidence it was unclear when assessments had been sat, rendering wider generalisations a challenge.

How did centres judge performance on assessments?

In order for students' work to be used as evidence to inform a TAG decision, teachers had to judge their students' performance in these pieces of work, and specifically in relation to the subject construct and grading criteria of the TAG for the particular qualification being graded. This section presents findings that gave us some insights into this process. We draw attention in particular to findings relating to i) the use of marks and grades by centres and ii) the use of grade boundaries.

Marking and grading

Within various centres' submissions, we found assessment results (i.e., marks or grades) that students had achieved on individual assessments used to inform final TAGs. We decided to inspect the use of marks and grades further within the data, as the way centres judged performance, especially whether this was done in relation to marks or grades, will likely have affected how centres combined this information to arrive at a final TAG and how cognitively demanding this task may have been. Marks and grades, to some extent, provide quantitatively and qualitatively different information about students' performance, and marks on different assessments can be less comparable to each other than grades. Thus, determining the final TAG may have been more demanding when having to combine information judged in different ways and when the assessments were only judged with regard to marks (Williamson, 2018).

Across all the centres we analysed, results data for individual assessments varied with regard to which types were found (marks or grades) and how these results were provided within the centres' submissions. Such variation was found both between centres as well as between assessments or candidates within the same centre. For English, several centres submitted a table confirming the individual assessment results, but some contained marks only, others contained grades only, and some recorded both marks and grades. Most of these centres noted the raw marks but a few recorded the mark as a percentage of the total marks instead. It was rare to find grades on individual assessments without accompanying marks. For the centres that did not provide such results data, we sometimes found marks or grades on the candidates' scripts themselves. For Maths, while a minority of centres also submitted tables of assessment results, many did not. Instead, assessment marks were usually found on the assessment scripts themselves, and it was far more common to find marks than grade information. For both subjects, nonetheless, there were various cases where we could not find any marks or grades for individual assessments in the evidence submitted by the centre. When there was a lack of assessment results, it was impossible to know which approach (i.e., marking or grading, or both) the centre had taken.

For several centres, within both subjects' data, we also found some details of the procedures they took for marking or grading assessments. The aspects we found evidence for were: double marking, script anonymisation, moderation, and use of grade boundaries and grade descriptors. This evidence was found in different places within the submission data; in some cases, details were found within centres' documentation and in other cases insights into

these procedures were found from inspecting the annotations on candidates' assessment scripts. A few centres provided documentation with specific notes about marking or grading processes they followed. For example, one Maths centre's documentation included specific details about the moderation procedures taken for each piece of assessment evidence, with phrases such as "moderation of random anonymised samples used to ensure correct marking" and "external moderation by 4 other schools" typical. In some academy chains, there was also evidence of cross-trust moderation of marking. One English centre also drew attention to their use of moderation as well as double marking and script anonymisation, although they did not explain how they implemented these aspects.

In other cases, the comments about marking procedures were embedded within other documentation; for one other English centre we found comments about using mark schemes within their rationale for their assessment choices: "using valid assessments by the way of past papers… utilising OCR approved resources and marking schemes." Inspecting the annotations noted on candidates' assessment scripts also revealed evidence of double marking and moderation. For example, in English, some scripts contained marks written in different coloured pens, suggesting it had been double marked, and, in the Maths sample, candidates' assessment scripts contained evidence of initial teachers' marks having been moderated both by other teachers within the centre and, in some cases, teachers from other centres, in order to ensure that the marking was accurate. Many centres highlighted in their documentation that this double marking was part of their moderation processes.

Furthermore, there were two differences that stood about between the Maths and English data with regard to the information we found about their marking and grading processes. One difference was the reference to grading methods. In the Maths' sample, many centres' documentation focused on their use of grade boundaries. Some English centres also mentioned using grade boundaries in some way to grade individual assessments, but that was much less commonly found. The next section discusses these findings on grade boundaries in more detail. Additionally, two English centres explicitly mentioned using grade descriptors in their documentation, but this was always discussed in reference to making judgements for the overall TAGs rather than for judging performance on individual assessments. There was no substantial evidence that Maths centres explicitly mentioned using grade descriptors.

The second difference between Maths and English related to the content of the marking commentaries on the candidates' scripts. It was common to find marking commentaries within the English scripts that referred to assessment criteria and grade descriptors. Many of these seemed to be there as justifications for the teachers' marks or grades. However, on various English scripts, some of the teachers' comments were directed to the student, suggesting this was not simply for their own marking/grading process. Some of these comments, for example, gave performance feedback to students on their responses, others asked students questions (e.g., did they run out of time?), and other comments seemed more motivational in nature. The process of giving comments to the students on their assessments could have had various effects on the marking or grading outcome. For example, teachers may have marked assessments more harshly, in order to support future student motivation, or conversely given the benefit of the doubt on the basis that formative assessments are part of a development process. There may have also been multiple uses for these assessments, whether complementary or contradictory. The judgements made by

teachers on assessments on which they give feedback may be cognitively somewhat different to those treated more summatively – they may be thinking about the process in a distinct manner. This factor is important to note, especially as there was variation both between and within centres with regard to whether student feedback was given on assessments. There was no evidence of these types of marking commentaries for Maths.

Setting and using grade boundaries

As mentioned above, the use of grade boundaries is one particular aspect of grading that some centres provided specific details about within their documentation, although this was more often commented on by the Maths centres than the English centres. Some centres indicated that they directly used previous grade boundaries set by OCR to determine their students' grades on individual assessments, that is, they used them in the way they were intended to be used when OCR set them. For example, there was evidence in Assessment Records that Maths centres mostly used OCR grade boundaries directly when grading full (as opposed to partial) exam papers; centres took the grade boundaries that were set when the papers were sat by the original cohort of candidates and identified students' grades based directly on where their marks fell relative to these boundaries.

However, there were also various examples of centres modifying OCR grade boundaries in both the Maths and English samples. Some centres explained that they modified grade boundaries because only part of the exam paper was used for the assessment, and they adjusted the boundaries directly in relation to the amount of the paper that formed the assessment they used for TAGs. For example, some Maths centres simply recalculated the percentage of the overall marks available that the original grade boundaries represented, and then compared candidate scores as percentages to these boundaries. Another Maths centre described that they had "averaged and scaled" (in a way further unspecified) to ensure "grade decisions were based on the marks available on the papers." One English Language centre explained that they halved the grade boundaries on full papers to obtain the boundaries for the two separate sections of the paper (Section A or B), as they were each worth half of the total paper marks. This centre acknowledged that "the boundary weighting might not be equal across the paper" given that "most students perform better in Section B" but, despite, this, they felt these grades were useful for assessing the grade at which students were performing. This potential difference in performance between Section A and Section B of the English Language exam was also highlighted by another centre. Because of this, that centre calculated the percentage of the overall marks that students achieved in each of the sections of a full paper that candidates had taken, and graded each section separately (while also providing an overall grade). However, not enough detail was provided to understand precisely how these sectional grade boundaries were derived.

In another example from English, one centre did not base grade boundaries on the boundaries set for the exam papers from which the questions came from. Instead, they chose to use the grade boundaries they felt represented the standard they wanted to work with. This particular centre used the 2018 grading structure as it was deemed to represent the level between 2019 and 2020. This centre even combined marks from two sections of two different papers to be able to obtain overall grades when only a partial paper was taken.

One Maths centre utilised assessments from the 2017, 2018 and 2019 sessions, but used the 2019 grade boundaries as a starting point for grading all of them, for reasons which are

not clear from the centre's Assessment Record. However, due to the fact that the assessments were sat at different times throughout the course, not all the subject content had been covered by the time two of the assessments were taken. The centre, thus, chose to reduce the grade boundaries. This was by six or seven marks in the case of one of the assessments. For the other, "the 2019 grade boundaries were amended to account for the fact that only 57% of the content on the paper had been covered by January of Year 10 when the students sat this paper." It appears from script evidence that candidates were invited to attempt all the questions, even when they had not yet learned the content; this is supported by the fact that all of the assessments were full papers of 100 marks. However, some elements of the Assessment Record appear to contradict this account, which suggested that the candidates were not assessed on content they had not learned. This is not the only discrepancy between different aspects of this centre's Assessment Record, and this, again, highlights the challenge for this research in being able to determine what exactly centres were doing, using only the information available to us.

Furthermore, there was not only variation in how the grade boundaries were used but there was some evidence of variation in the purpose of these grade boundaries. For example, one centre in the English sample specifically explained that their grade boundaries were used for "sense-checking historic thresholds in light of grade descriptors and level of control".

Overall, it can be seen from these examples that, while some centres placed some value on the use of grade boundaries as a way of understanding the ability of their candidates, some used them in ways dissimilar to the ways they are used in normal years. Modifications to boundaries, or the use of boundaries in non-standard ways, were common, and again there was significant variation between methods utilised by centres.

How did centres determine the final TAGs?

All the centres we reviewed used multiple pieces of evidence to determine TAGs for every (or almost every) student for GCSE Mathematics and English Language; it was only in a few cases that only a single piece of evidence was used, due to candidates being unavailable for the other assessments. This meant that all centres had to combine different pieces of evidence in some way to decide on the final TAGs. Thus, this section focuses on the information we found about how centres decided on final TAGs, and their rationales. For both subjects, many centres did not provide information as to exactly how they determined final TAGs from the set of evidence they gathered. However, there were several centres for each subject that provided details of their processes, which gave us some useful insights into the approach they took and how they prioritised evidence, and revealed some interesting similarities and differences between centres.

Prioritising particular pieces of evidence

Several centres in both the Maths and English samples made comments that explained, or strongly suggested, that they placed higher value on certain pieces of evidence than others when determining their students' final TAGs. In both subjects, there was evidence that, across the centres, these views were affected either (1) by characteristics of the assessment or (2) by circumstances around the individual students, including their performance levels, or by both assessment and student factors.

1. Characteristics of the assessments

Three types of assessment characteristics were mentioned within centres' commentaries on why they used certain assessments' results more or less when determining TAGs:

- the type of assessment evidence,
- the assessment conditions, and
- the date the assessments were sat.

Various centres, in both subjects, drew distinctions between different types of assessments, specifically distinguishing between full exam papers and exam-style assessments. One Maths centre explicitly stated that they valued full exam papers more than exam-style assessments: they explained that they used a session's worth of full papers as their primary evidence (their sampled candidates had taken between two and five full papers) and a set of smaller exam-style assessments (between two and four pieces per sampled candidate) served as lower priority evidence. The greatest priority was given to the full papers, according to multiple Maths centres, because they covered all the content and AOs.

In contrast, for English, there were differences between centres with regard to their relative use of full and exam-style assessments. Some centres seemed to align, to some extent, with the Maths findings, placing particular weight on full papers. For example, one English centre explained that they treated performance on whole papers (as opposed to performance on sections of a paper), as the "benchmark" for deciding TAGs, which seemed to be because it gave information about students' overall ability in the subject. However, this centre, as well as many other English centres, also disaggregated performance on these full papers when considering TAGs, specifically so they could consider candidates' performance on the two distinct sections of the full papers separately (Section A – reading and Section B – writing). That centre explained that there was a limitation of using overall performance on the full paper to make their final TAG decision, which was that most of their students performed differently in the two sections of the paper. Another centre that also considered assessments of the two content areas separately when deciding TAGs explicitly attributed this to JCQ guidance: "JCQ separated the grade descriptors into [1] Critical Reading & Comprehension and [2] Writing". Therefore, for some English centres, the different types of assessment (full and exam-style assessments) seemed to serve different purposes in supporting their TAG decisions rather than one type being favoured over the other.

The conditions under which assessments were taken also seemed to affect how evidence was prioritised in reaching final TAGs – with several centres explicitly mentioning this. Maths centres tended to highlight that the best evidence they had was that which was collected under full or relatively formal exam conditions, which some said meant these assessments were the best for determining their candidates' individual ability. In the English sample, one centre similarly explained that they placed greater weight on assessments with higher levels of control. In most of these examples, however, it was not clear which aspects of the exams or high level of control centres placed particular value on (e.g., silence, no communication between candidates etc.). For example, one other English centre explained that one assessment was "important data" for their TAG decisions but specifically because the exam paper was unseen by candidates. This same centre was one of the few centres that gave very detailed TAG rationales for their individual candidates, and throughout these

commentaries they often highlighted the level of performance demonstrated under “timed conditions” in comparison to performance in classwork or homework. This centre appeared to value evidence of candidates performing at the TAG grade in timed conditions, and seemed to weight performance in these timed assessments more highly when there was a discrepancy with untimed performance.

The times of year when assessments were sat also appeared to factor into some centres’ prioritisation of evidence. Many Maths centres used only material from April and May 2021, but, in cases where earlier material was used, it was sometimes explicitly given lower priority in TAG decisions or used only where there was a paucity of candidate evidence from April and May 2021. One Maths centre highlighted that this was an explicit decision that was made due to the fact that less content had been covered earlier in the assessment cycle, and that the “demand” of the most recent assessments was higher as they assessed everything. Similarly, one English centre also explained that they placed greater weight on assessments that were completed most recently (i.e., in the summer term of 2021).

2. Individual students’ circumstances and performance

Within centres’ rationales for individual students’ TAGs, we found some comments about students’ circumstances or their performance levels on particular assessments, which seemed to affect how the teachers weighed up the students’ different assessment results. This meant that the same pieces of evidence could have been weighted differently for different students, even within the same centre.

Some centres described taking into account aspects of the circumstances surrounding the student at the time of the assessment. Various English centres mentioned placing less emphasis on certain pieces of evidence where there had been mitigating circumstances around the assessment, including a bereavement, a medical condition and a family break-up. There was less evidence in the Maths context of explicit reference to mitigating circumstances as a reason for giving evidence a lower priority but still including it; more typical was assessments not being used at all. Another English centre drew attention to the lack of access arrangements in place for one student for some assessments, which they explained meant that the performance in those assessments was a poorer indication of the student’s “true ability”. There was, however, variation in the level of explanation centres gave for making different prioritisations for different students; many comments were vague. To give an extreme example, one English centre outlined which assessments were the “best evidence” for that candidate, but they gave no reasons why they judged them to be “best”.

Within the English sample specifically, there was one centre where teachers provided a very detailed account of their TAG rationales, which revealed a much wider range of student factors being considered in their quest to understand why individual students performed differently on different assessments. This centre provided detailed explanations of all discrepancies between assessment results; for example, one students’ diverse results were attributed to the student performing worse in timed than untimed assessments and another students’ increased performance in more recent assessments was attributed to improvements in their exam technique and timing.

A different example of centres taking into account student factors was when TAG decisions were affected by how well students performed on the assessments, seemingly irrespective

of any other conditions around the student or the assessment. For example, one of the English centres seemed to simply base their TAGs on whichever of two assessment results (on two full exam papers) was higher. This centre acknowledged that the better outcome was usually the first attempt, but they did not provide further reasoning for not taking into account the results of both assessments, and why the students may have performed differently in the two assessments. In Maths, there was also evidence of a 'best grade' approach being used by at least one centre. This centre set their students a full set of six half-papers in April 2021 and a similar whole-specification set of exams in November 2020. In their documentation, they explained that the higher-scoring of the two sessions' results would be used, though in the event the session sat in April was the only one from which evidence was provided for any of the sampled candidates, implying that these candidates (at least) scored higher on these assessments than those they had sat earlier in the year. It is not known whether, for any of the candidates at the centre not included in the sample, the earlier session was the higher-scoring, and, thus, determined their TAG.

Variations in approach to combining information

Whichever way pieces of evidence were prioritised, the different assessment results needed to feed into some kind of process where the teachers combined this information to arrive at an overall TAG. The JCQ guidance on determining TAGs repeatedly mentioned the importance of taking a holistic approach to grading, and emphasised that, this year, grading was "not intended to be a formulaic calculation, and should account for the context in which each student's evidence has been produced" (JCQ, 2021, p.24). However, there are various ways to approach the idea of holistic grading, and, with centres given the freedom to define their own processes but also working under resource, time, and other pressures, how this was operationalised was likely to have differed significantly across centres. This section presents findings that gave us insights into the final combinatory process that our sampled centres took to determine their students' final TAGs. These findings predominantly come from the English sample, because we could not find explicit information in any of the Maths centres' documentation explaining the whole of this process. Hence, for Maths we could not objectively identify if the final TAGs were reached in a formulaic way (e.g., based on the modal or mean grade) or a less tangible method of coming to a holistic judgement.

For English, the notion of taking a holistic approach to TAGs was a clear common theme throughout the sample reviewed. The terms 'holistic' and 'holistically' were written explicitly within many centres' documentation, sometimes specifically in reference to the centres' approach to combining information for final TAGs. In fact, every one of the centres that mentioned using previous grade boundaries for individual assessments emphasised that the final TAGs were determined holistically. One centre even noted explicitly that they did not apply a calculation to the assessment grades due to JCQ regulation. Individual assessment grades thus seemed, in some cases, to be used to help centres evaluate and determine their students' overall grades rather than being used for numerical aggregation purposes. However, most English centres did not go into detail about what precisely this holistic process involved in practice, neither in terms of their general strategy nor for individual candidates, so we often could not determine how this occurred and whether some kind of more formulaic process had been employed. The same was true for the large majority of Maths centres with regard to the paucity of specific information on how centres drew together the evidence from different assessments into one final grade.

For a few centres across English and Maths, nevertheless, we found a level of detail that gave us some specific insights into the TAG decision process and revealed various similarities and differences in approaches taken between centres. For English, two centres provided information about both their general strategy to form holistic TAG judgements as well as separate rationales for individual candidates' TAGs. These two centres were similar in that they both stated they took a 'best fit' approach to determining their students' final grades, although, in practice, this did not seem to equate to exactly the same approach.

The first English centre explained that 'best fit' meant they were looking "for most consistent trends in their [students'] performance", and this seemed to be both for performance between assessments testing the same content area (e.g., two assessments on writing) and for performance between content areas per the JCQ's distinction between reading and writing within the grade descriptors. The TAG rationale this centre submitted for every sampled student was very detailed; each contained a detailed explanation for (1) their "holistic judgement" of the student's overall TAG (e.g., grade 6), (2) "holistic judgement" of their grade for Critical Reading & Comprehension (e.g., grade 5), and (3) "holistic judgement" of Writing grade (e.g., grade 7). Within this centre's overall summary of their TAG approach, they explained that the overall grades lay "in most cases" between the grades of the two content areas but there was no clear evidence in any of the candidates' TAG rationales suggesting a mathematical approach was used to determine the final grade (e.g., the grades were not averaged). Sometimes there were references to aggregated mark percentages across certain assessments, but aggregation of this kind appeared to be mainly used to provide indications of the student's overall levels of performance in the subject. Instead, the TAG rationales showed the centre considering a range of factors qualitatively, including the different skills stated in the grade descriptors, how well and how consistently the candidates had demonstrated them, the conditions under which assessments were taken, discrepancies between assessment results, and candidates' progression of learning or performance. Some of the commentaries even included what candidates would have had to have demonstrated to have been given the next grade up.

The second English centre that also mentioned taking a 'best fit' approach seemed to focus more specifically on the relative performance between the two content areas of reading and writing. The centre went into some detail about how they used "section grades to inform a 'best fit' approach", but the TAG rationales for individual students revealed differences between how this was implemented across students. For example, for one student, the teacher's commentary seemed to describe a formulaic approach to 'best fit'; they explained that section marks were added, averaged and then doubled to get to an overall grade. However, for another student in this centre, there did not seem to be a mathematical formula involved, especially not one based on aggregating marks, but was based on a more qualitative evaluation of the balance between reading and writing skills: "He excels in Section B Paper 2 questions, but is less strong in Section A. Therefore, although he consistently shows evidence of [grade] 9 in Section B, I feel a grade 8 is more appropriate as it is an accurate representation of the breadth of skills across components."

It is important to note that not every centre had to combine the results from different assessments at all. One English Language centre based their TAG decisions on identifying the single best evidence of their students' ability rather than on combining different pieces of assessment results. This centre produced TAGs from the best attempt at two full exam

papers. A small number of Maths centres derived grades based largely on taking the marks achieved for a whole set of assessments that added up to something in the vicinity of a whole qualification's amount of assessment and then comparing these total marks to pre-existing grade boundaries. In other words, there was effectively a single grade representing the whole qualification. This was only the case for the small number of centres where assessments were chosen to represent the equivalent of exactly (or almost exactly) one full qualification. Of course, in all these cases, the overall grade can still be viewed as holistic, because these assessments were holistic in nature in that they assessed content from across the range of subject content; it is just that there was no specific role for teachers in making a holistic judgement and determining the overall grade themselves.

Quality assurance procedures

We found a couple of pieces of information about some centres' procedures for quality assuring the overall TAGs. Unlike in other sections, it was not surprising to have found only small amounts of information on this because these procedures were intended to have been captured in Centre Policy documents, as centres were expected to take a consistent approach across all subject departments. Nevertheless, some Heads of Departments in Maths and English mentioned these procedures, and this gave some insights into how these kinds of procedures were implemented in these specific departments.

One English centre explained that they undertook a moderation procedure led by the Head of Department, as follows: "set teachers wrote full holistic judgements aligned with the grade descriptors for each separate pupil, which allowed the HOD [Head of Department] to moderate and standardise accordingly across the Department." Another English centre provided a document containing notes from a "Quality Assurance meeting", where two teachers appeared to discuss the TAGs of candidates and to agree on the grades. For example, notes included statements such as: "TAG for [student name] discussed… It was decided that there was enough to award a TAG of grade 3", and "[the teacher] shared grades awarded and explained how they had been awarded". In Maths, some centres suggested that external and internal moderation as well as Head of Department quality assurance also took place in relatively similar ways to in English. However, it was more common for Maths centres to provide evidence about the ways they had quality assured their marking than their final TAGs. This connects to the sense that TAGs (both at individual assessment level and, in some cases, the final TAG) were awarded in Maths in ways more akin to normal procedures of comparing awarded marks to grade boundaries, and as such that it was more important for centres to demonstrate that the marks were correctly awarded.

Discussion

The aim of this research was to gain insights into how centres determined TAGs in 2021 and what evidence they used to do so, with a focus on variations in approaches. Previous research had given us some understanding of the types of evidence used, by asking teachers to give accounts of their process via surveys (Holt-White & Cullinane, 2021; Johnson & Coleman, 2021), interviews (Johnson & Coleman, 2021) and diaries (Johnson & Coleman, 2021). This current study took a different approach; we inspected the actual evidence submitted by centres to OCR as part of its TAG quality assurance processes, and we examined this data with regard to both what evidence was used as well as how centres made their TAG judgements. Our analyses focused on the submissions for GCSE English Language and GCSE Mathematics of the centres sampled by OCR.

How useful was the data for understanding TAG processes?

A key unknown at the start of this project was how useful the centres' submission data would be for our research aims – to understand and draw comparisons between centres' TAG processes. This was because centres' submissions were intended for external quality assurance, specifically, to enable quality assurers (from awarding organisations) to decide whether or not the TAGs of individual centres were based on reasonable evidence and judgement. Our analyses found that there was a rich amount of information in each centre's submissions, which provided us with various important insights into: the evidence used (types, origins, subject content coverage, amount); the conditions under which the assessments were taken (level of control, duration, dates); how centres judged performed on assessments (marking, grading); and how centres determined the final TAGs (prioritisations, approach to combining information, internal quality assurance).

The data, however, also had several characteristics that constrained the analyses we could conduct and limited the conclusions we could draw. First, there were data inconsistencies both between and within centres' submissions (e.g., how centres recorded and submitted evidence, and how centres described and differentiated pieces of evidence). Second, when information was provided, it was sometimes partial or vague (e.g., with regard to the specific details of the assessment evidence and TAG rationales). Third, in some cases, information about certain aspects was missing completely (e.g., we saw many assessments without information about the date they were taken, or the origin of the questions, and there were also many submissions without TAG rationales). Together, these features meant we could not analyse the data quantitatively and therefore were unable to provide precise answers to questions about the level of variation (e.g., how many centres used the same types of evidence?). This set of features also restricted our qualitative analyses, to some extent. In particular, as every centre did not use the same submission approach, we could not draw any conclusions from any absences of information, which in turn inevitably limited our capability to compare and contrast centres.

Overall, despite these data limitations, the information within centres' submissions enabled us to increase our understanding of the TAG process and the variations that occurred both between and within centres, adding to the findings of previous research (Holt-White & Cullinane, 2021; Johnson & Coleman, 2021).

What did the centre submission data show?

Table 5 and Table 6 provide summaries of the findings. Table 5 focuses on features of the assessment evidence and Table 6 focuses on how the evidence was used. In Table 5 we have contrasted the findings we frequently observed against the wider range of variations (this distinction was not possible for Table 6 as the data was less comprehensive).

Table 5. Features of the assessment evidence found within the submissions.

	Frequently observed (no features were unanimous)	**Variations observed (between or within centres)**
Assessment evidence		
Type	• Exam-style assessments used • Combination of different exam-style assessments (e.g., full paper, half paper, single exam question)	• Exclusive use of full exam papers • No use of full exam papers • Non-exam-style evidence (classwork, homework quizzes) used
Origin	• Use of OCR GCSE materials • Specific use of 2019 and 2020 GCSE materials, including Additional Assessment Materials	• OCR 2017, 2018 or legacy GCSE materials used • AQA or Pearson materials used • Materials not from awarding organisations (e.g., textbooks, Maths websites) used
Content coverage	• Candidates tested on broad subject content, covering all AOs covered in GCSE exams • GCSE Speaking AOs not assessed (English only)	• Exclusion of content not taught from assessments • Inclusion in assessments of content not taught • More assessment of certain content areas (e.g., writing in English)
Amount	• Multiple assessments used • Approximately same volume of assessment as in a normal GCSE session	• Much more assessment than in a GCSE session • Much less than in GCSE session • Much more of certain content than in a GCSE session (e.g., writing in English)
Assessment conditions		
Centre-defined 'level of control'	• Assessments described as taken under “exam”, “formal” or “controlled” conditions	• Assessment conditions defined as “high” control • Assessment conditions defined as “medium” control • Assessment conditions defined as “low” control
Specific conditions under which assessment taken	*[Not enough data]*	• Assessments taken in exam hall (Maths only) • Assessments taken in classroom • Assessments taken in other locations (e.g., home) • Combination of assessment locations • Invigilated or supervised assessments • Open book assessments (English only) • Closed book assessments (English only)
Duration	*[Not enough data]*	• Timed assessments • Timings aligned with normal GCSE exam timings • GCSE timings extended by centres due to students’ lack of exam experience • Timings decided by centres • Untimed assessments
Date	• At least some assessments taken in April or May 2021	• All from April/May 2021 only • None from April/May 2021 • Assessments from 2018-2021 used • All taken from a small time period • Taken from a wide time period

Table 6. Variations in centres' judgemental processes found within the submissions.

	Variations observed (between or within centres)
Assessment judgements	
Marking and/or grading of candidate performance on assessments	• Marks and grades provided • Only marks or only grades provided • Neither provided
Marking or grading procedures	• Double marking • Script anonymisation • Within-centre and cross-Trust moderation of marking procedures • Use of mark schemes • Use of grade descriptors (English only) • Marking annotations on some scripts • Some student feedback (English only)
Which grade boundaries used	• Those from session from which exam paper originated • 2019 grade boundaries (centres' choice of standard) used for non-2019 papers • Centre-modified boundaries to account for reduced content • Centre-derived boundaries when they did not previously exist (e.g., for sub-sections of the paper) (English only)
Purpose of grade boundaries	• To determine final TAGs • For sense-checking
Final TAG judgements	

Prioritisation of evidence	• Based on assessment characteristics
	  ○ Prioritising overall results on full papers
	  ○ Distinguishing between results in different content areas (i.e., writing vs reading)
	  ○ Prioritising evidence taken under higher levels of control (e.g., exam conditions, timed, unseen papers)
	  ○ Prioritising more recent evidence
	• Based on individual students' circumstances or performance
	  ○ Reduced emphasis on evidence where students had mitigating circumstances or inadequate access arrangements
	  ○ Exclusively basing TAGs on students' highest performance or set of best grades

Combining performance information	• Explicit mentions of "holistic" approach to determining TAGs (English only) • "Best fit" approach being taken to combine results of different assessments (Maths and English), or results in different content areas (English only) • Evaluating consistency of performance across different assessments (Maths and English), or different content areas (English only) • Diverse factors being taken into account during the TAG decision (e.g., grade descriptors, consistency of performance, assessment conditions, discrepancies between results and progression of learning) • No combination needed (e.g., TAGs based on results of one full paper or on one grade derived from a set of marks akin to qualification-level grade)
Internal quality assurance	• Head of Department-led moderation • Internal or external quality assurance meetings

How did the TAG evidence compare to that of typical GCSE sessions?

Although TAG policy and official guidance permitted centres flexibility to decide their own approach, they also contained recommendations that often aligned with approaches taken in a typical GCSE exam. For this reason, it is interesting to consider the findings in relation to the typical GCSE session. The assessment evidence we found in the submission samples had both similarities and differences to the assessments used in typical GCSE sessions.

Similarities to the typical GCSE session

Almost all of the features of the assessment evidence that we frequently observed (Table 5) reflected those of the typical GCSE exam session in some way or another. These were:

- use of exam-style assessments,
- assessments based on (or resembling) GCSE exam paper materials from awarding organisations,
- candidates being tested on all GCSE AOs,
- multiple assessments being used, amounting to approximately the same volume of assessment as in a GCSE session,
- most assessments being taken under exam-like conditions, and
- at least some assessments taken late in the academic term (April/May 2021), similar to when GCSE exams would have been held in a normal session.

Although none of these features were unanimous across the samples we analysed (i.e., true for every candidate in every centre sampled), the high level of occurrence suggests that the ordinary structure of the GCSE exam session held importance for teachers. This coheres with previous research on TAGs, which also reported various similarities between TAG assessments that teachers used and GCSE assessments (Holt-White & Cullinane, 2021; Johnson & Coleman, 2021).

For most centres we could not determine from their submissions the specific reasons that they incorporated these features in their TAG approaches. However, a variety of benefits were mentioned by some centres when justifying their assessment choices, including that:

- official assessment materials were designed to cover the required AOs,
- assessment materials and the marking of them could be standardised,
- teachers could put in place access arrangements for students, and
- the exam experience could be replicated for students as closely as possible.

There are, of course, other possible reasons why these features may have been common across centres. For example, Ofqual and JCQ recommended many of these aspects in their TAG guidance to centres, which centres could have viewed as particularly important or more justifiable when it would come to quality assurance. Combined with this, teachers' understanding of, and familiarity with, the assessment structure of typical GCSE sessions could have formed a schema (an organising framework) for them to make sense of the wide ranging information being presented to them in the guidance (see McVee et al., 2005 for a discussion of schema theory). Another explanation is the possibility of some of these

features having less detrimental effects on teacher workload, which had been increased by the fact that teachers were required to determine their students' grades and by the wider effects of the pandemic on teaching and learning (Johnson & Coleman, 2021). The use of exam-style assessments based on GCSE materials, for example, could have reduced teachers' workload, as they would not have needed to spend time designing their own. It could also have reduced uncertainty around whether they were selecting the 'right' kinds of evidence (Johnson & Coleman, 2021).

In addition, there were other features of assessments that were observed in some centres' submissions that also reflected aspects of the typical GCSE session. For example, some assessments were full, unmodified, copies of GCSE exam papers, taken in exam halls, taken under timed conditions to the duration typical of a normal session and previously unseen by students. These aspects were not specifically recommended in official guidance, which suggests that the value that some teachers placed on the GCSE format was not necessarily linked to TAG requirements or compliance. Again, most centres did not provide explanations for incorporating these features, and therefore we could not determine which advantages they felt these assessment features had. The few direct comments we found were about full papers acting as benchmarks of students' overall performance and the unseen nature of exam papers making assessment results "important data" to a centre.

Deviations from the typical GCSE session

The sample of assessment evidence we analysed also contained many features that differed to that of typical GCSE sessions. Deviations were expected, given that it was a policy intention for centres to be able to collect evidence of students' attainment in a way that fitted around the circumstances of their students during the pandemic. Some deviations seemed to align with recommendations JCQ set out to ensure teachers could determine the appropriate TAGs for their students, such as modifications of GCSE exam papers to remove non-taught content and the rare use of non-exam-style evidence, specifically for when exam evidence was not available. Other deviations also seemed to relate to the pandemic, affecting, for example, where assessments were taken (e.g., remotely, at home) and sometimes when assessments were taken by students.

A range of other differences from typical GCSE exams did not seem to link to specific TAG recommendations or the situation of the pandemic, and it was, in many cases, not clear why some centres decided to incorporate these features in their assessment approaches rather than replicating the typical GCSE process more closely. For example, teachers often used combinations of different exam-style assessments (e.g., full papers and half-paper; full papers and single exam questions) rather than simply using full exam papers as in a typical GCSE session, while other centres did not use any full exam papers. The decision to use partial papers may have been practical (and pandemic-related). One Maths centre explained they used a set of half papers so they could be taken during class time across several classes. Practical factors like these might also explain other deviations such as assessments being taken under teacher supervision, assessments being completed as homework rather than in class, perhaps so to not reduce teaching time, and assessments being completed across academic years rather than at the end of the GCSE.

A different type of deviation from GCSE, which highlights some other potential factors at play in the TAG process, were the amounts of assessment used to inform TAG decisions in

centres. For some candidates at certain centres, there was considerably more evidence (in terms of marks, duration and content coverage) than in normal years, though it may have been distributed differently (i.e., across the academic year rather than all being taken at the end of the GCSE course). By contrast, other candidates had very little evidence, sometimes only a couple of small pieces. Many candidates fell in between these extremes.

There are many potential reasons why the amounts of assessment varied both between and within centres. Covid-19 is, again, one factor that may have affected how much assessment could be completed, both in terms of candidate availability but also indirectly due to how much class time could be devoted to assessment rather than teaching and learning. However, there are a range of other factors. For example, over-assessment (relative to a normal GCSE session) may have been affected by accountability pressures or by teachers wanting a high amount of evidence to be confident about their TAG judgement (Johnson & Coleman, 2021). Another possibility is that this volume could have been available in certain centres where assessment is a core part of the teaching and learning process. This may explain why in English, in particular, there were lots of single writing questions submitted as evidence; perhaps writing tasks are more routinely conducted in English classes rather than having been put in place to support teachers' TAG decisions.

What can we learn from the findings about teacher assessment?

Effects of variation in assessment approaches

The findings showed a wide variation in approaches between centres, which raises questions with regard to comparability of standards. This situation is very different to that which pertains in a normal GCSE session, in which all candidates in a subject (especially ones that do not contain optionality) take the same assessments under the same conditions and covering the same content. These are marked in the same way for every candidate, and their qualification grades determined using the same approach to boundary setting and the aggregation of marks from different assessments (i.e., different components). Given the variation we saw in this sample, it is fair to say that each centre was in effect operating its own way of evaluating whether the standards of GCSE grades had been met by their students, which inevitably poses questions for any claims about comparability between grades in 2021 that might be made.

In addition, there was evidence of variation at different levels too: between different candidates in the same centre, between different centres in the same subject, and between different subjects (i.e., Maths and English Language). The difference found between candidates within the same centre is perhaps the most concerning, given that this was, arguably, the most straightforward type of comparability to ensure at a centre (though, of course, there is a limit to how much centres could control for their candidates' distinct circumstances). Indeed, the official TAG guidance emphasised cohort consistency, requiring that, where possible, all candidates should be treated the same way in terms of the evidence used to inform their TAGs. Although some variation between candidates was inevitable given the differential and unpredictable effects of Covid-19 on students, in some cases this variation was more substantial than might be expected. For example, within one Maths centre, one candidate had nine separate pieces of evidence going towards their TAG, and another only three. There are many other examples of similar differences. It could be

suggested that this is a feature of candidate availability to take particular assessments, and that, therefore, the potential for different amounts of assessment could be justified as ensuring that there were enough opportunities for candidates to produce evidence, even if they were frequently absent for reasons of illness, bereavement or other issues. It is notable, however, that the official guidance did not suggest a maximum (or indeed a minimum) amount of evidence that should be considered, and as such this variation was allowed. Perhaps this ought not to be seen as a significant issue, in that the purpose of the guidance was to ensure that teachers were making holistic judgements of a candidate's ability, and such judgements may be made using both small and large amounts of material. As we highlighted in the introduction, the guidance was designed (in a context where those writing it were mindful of major discrepancies between centres in terms of how much teaching their students had received and how long they had been in the classroom) to be maximally flexible, ensuring that as many candidates as possible could receive grades.

A challenge for comparability worth drawing attention to is the difficulties that variation creates for any assessment (such as of performance standards) that is based on judgements using comparison. This includes using the Comparative Judgement approach as a method for comparing the standards of assessments between years or the standards of candidate performances against one another. However, it also relates to other situations in which judgemental comparisons are made, including in the script scrutiny phase of current standard maintaining procedures, or in the judgement of portfolios, for example. It is typically suggested (e.g., by Leech & Chambers, 2022) that comparisons of assessment material are more straightforward when the material being judged is comparable in structure (e.g., in terms of length, number of items, item types, and so on) so that the only differences between versions are differences between candidates' performances.

In a future teacher assessment context in which there were significant variations in terms of the kinds of materials used to support grading, the absence of this comparability in structure is likely to lead to more difficulty for those having to judge standards between centres (e.g., by moderators or other external quality assurers), especially if this was done in a strongly judgemental way, as opposed to a more statistical approach. The same may also be true within centres if teachers had to compare their students' performances against each other on the basis of differently structured assessments or sets of evidence that are dissimilar to each other. This all highlights the comparability benefits of consistency in assessment.

Effects of differences in understanding of assessment concepts

One prerequisite for common assessment approaches between centres is ensuring a common understanding of the assessment concepts underlying them. In our analysis of centres' submissions, there was some evidence that centres differed in their interpretation of the assessment concepts outlined in official TAG policy documents. For example, there were differences in how centres defined the "level of control" of their assessments (e.g., what a 'high' level of control meant), and in how they referred to types of evidence (e.g., mock exams or in-class assessments). Centres also differed in their interpretation of what constituted a 'single' piece of evidence; for example, some treated an assessment that comprised of two sections as one piece of evidence while others treated each section separately and listed them as two pieces. This finding highlights that some caution may be needed when interpreting the findings of studies that have reported the number of assessments teachers used to determine their TAGs (e.g., Holt-White & Cullinane, 2021). In

our context, it meant we were unable to directly compare centres to one another in terms of the number of pieces of evidence they used, meaning that comparability of assessment between centres had to be considered in a less quantifiable way.

The use of grade boundaries by centres was another finding that revealed that some centres had a firm understanding of the appropriate uses of this technical aspect of assessment (i.e., what grade boundaries can and cannot tell you) while others seemed less knowledgeable, as they used them in less appropriate ways. For example, some centres used the grade boundaries designed for one year's assessment to grade another year's paper, perhaps revealing a lack of understanding of the ways in which grade boundaries are set in order to account for differences in difficulty.

Centres also seemed to differ in their understanding of what a 'holistic judgement' was, in the context of drawing together varying pieces of evidence to form a single TAG, as there was wide variation in methods used to combine information to determine final grades. Some centres explained that they took a best fit approach to assessment results, but others seemed to focus on overall marks on a set of assessments, using grade boundaries to determine the grades. Both of these approaches can be viewed as holistic judgement but for different reasons. The former approach – combining (in a non-algorithmic fashion) results from different assessments – aligns more directly with the official TAG guidance on holistic judgement and requires the teacher to make the holistic judgement themselves. In contrast, the latter approach – grading based on mark aggregation across assessments – may also be considered holistic if the marks have come from assessments that have tested a wide range of content. The one example we found that seemed to go against the spirit of holistic judgement was where centres reported that they focused only on the best grades achieved by candidates across the material presented for them. Although JCQ provided worked examples of the holistic judgement process, these findings suggest they may not have been sufficient for centres to be able to consistently apply them, without also being accompanied by a specific definition of the concept of 'holistic judgement'.

This issue of centres having different understandings of assessment concepts may have emerged because definitions of these concepts were not provided within any official guidance on TAGs. According to cognitive psychology research (e.g., schema theory; McVee et al., 2005), when information is presented without a clear conceptual framework, it creates the condition for readers' prior knowledge to strongly influence their interpretation of the information. In the case of TAGs, teachers' prior experience with assessment concepts may have affected their interpretation of these concepts when mentioned in the TAG guidance). Thus, a key conclusion, and an important consideration for the future uses of teacher assessment, is that we cannot assume that centres share a common understanding of assessment types, assessment conditions, assessment processes or assessment terminology, without these being defined specifically for them. Guidance should contain a conceptual framework to guide teachers' interpretations to ensure the concepts within it are understood consistently, and as intended.

Validity and reliability of teacher assessment

Teacher assessment is widely acknowledged as providing the opportunity for more valid assessments in certain contexts. The argument is that teachers have access to a rich range of information on students' level of understanding of subject content (Johnson, 2013; Vitello

& Williamson, 2017). In the submissions we analysed there was some evidence of teachers contextualising students' performance on assessments in a way that the typical exam process does not, which may have increased the validity of TAGs. For example, one centre provided detailed TAG rationales for their students in which they analysed the discrepancies between each student's results on different assessments to determine why performance was better or worse in some assessments than others, and, accordingly, which results were considered most indicative of their current ability in the subject. In the typical exam process, all terminal assessment results contribute to the final grade and no allowance (except for special consideration) is made for performance differences on different assessments.

However, other findings raised questions around the validity of the assessments being used as the basis for TAGs. For example, we found that TAG evidence, in some cases, was drawn from throughout the GCSE course, came from assessments or classwork that may not have been considered summative at the time it was sat, and may not have covered all the content. In an exam session, all the assessments are taken at the same time at the end of the course, and should, theoretically, therefore be assessing students at their optimal time of understanding the content and preparation for the exam.

It is important to consider whether the assessments that were used for TAGs were originally meant to be used in formative ways, from both the student and the teacher perspective, as in both cases validity may be affected, in different ways. We found no statements from centres as to whether assessments used for TAGs were specifically designed and taken as summative assessments or were taken for formative purposes and later used to inform TAGs. From a student perspective, this could have affected how they viewed the assessments (i.e., they may have tried less hard at assessments they saw as formative at the time they sat them, resulting in their ultimately final grade being based on material on which they were less motivated to do well) (Gill, 2019; Holmes et al., 2021). Teachers may also have given more or less support to students on assessments that they viewed as formative. In addition, there was evidence within English submissions that some teachers gave student feedback during the marking processes, which may have had different effects on their judgement-making (Delaney, 2005).

Finally, another set of questions that is raised by these findings concerns how much evidence teachers need in order to produce reliable grades via the process of holistic judgement. To what extent is this likely to be different from the amount of assessment normally used for GCSEs? Many centres in our sample used considerably more assessment than would have been the case in a typical GCSE session for these qualifications. On the other hand, there were also candidates in the sample whose teachers determined their grades from far more limited evidence. What constitutes enough evidence to justify a GCSE grade, even if a grade 1, in different assessment contexts is worth considering further.

What questions were raised by our research but left unanswered?

Our research raised many questions we could not answer with the data we had available, and which would be promising avenues for future research if appropriate evidence could be gathered. Some of these questions had already arisen when we were planning for this project (such as questions of bias, and the "why" of centres' processes) but the data turned out to be too limited, rendering it impossible to investigate them. Other questions arose

during our analyses. We have presented some of the questions below, dividing this list into ones that are specific to the 2021 process and those relating more generally to teacher assessment.

On centres' TAG processes in 2021

- To what extent were assessments designed to be taken specifically for the TAG process, as opposed to TAGs being derived from materials designed for other purposes, such as ongoing/formative assessment?
- What factors affected teachers' choices of assessments and the conditions under which they were to be taken?
- What did teachers understand "holistic judgement" to mean?
- How did centres decide on the approaches to take in their Centre Policies?
- Did centres have more evidence than they used for some candidates, but presented only the most favourable evidence to support the TAGs they awarded?

On teacher assessment more generally

- How does whether an assessment is perceived as summative or formative affect student performance on it and/or teachers' judgements of that performance?
- Does giving feedback on students' work affect validity of marking?
- To what extent, and via what specific processes, are teacher-awarded marks/grades vulnerable to unconscious bias? How can these biases be understood and addressed?

Recommendations

As a consequence of this research, we make the following four recommendations for improving possible future teacher assessment processes in order to enhance consistency, efficiency and comparability of standards:

Recommendation 1

In any future situation in which grades are to be awarded by teachers, centres should be required to provide **information in more consistent, more easily analysable formats**: for example, information (such as on specific aspects of centre policies) could be provided in spreadsheets or as responses to online forms.

This is not only because this would support researchers in terms of understanding the data, but also because consistency of format requirements is likely to lead to more consistency of content, which would ensure that centres' processes converge on a smaller number of approaches.

Recommendation 2

In future situations in which centres must gather evidence to support grading, guidance on which evidence to use must be as **clear and as explicit as possible for centres.** This should take into account the influence of teachers' prior experiences and knowledge of assessment concepts. Guidance should include:

- an explicit statement on the (maximum and minimum) amount of evidence needed, to reduce the likelihood of under- or over-assessment
- a hierarchical list of acceptable assessment alternatives, which includes the precise conditions under which each alternative is allowed, and
- clear definitions and explanations of assessment concepts and their appropriate use, including for grade boundaries and levels of control on assessments.

Ofqual's centre guidance for 2022 contingency arrangements for general qualifications has already moved in the direction of more specificity, especially with regard to assessment conditions and the volume of assessment (Ofqual, 2021a).

Recommendation 3

Should centres be required to provide TAGs in future, specific guidance should be given to support them to provide **detailed explanations of exactly how they draw together assessment evidence they used into a final grade**. This should include:

- discussion of how grade boundaries were used,
- weighing up and prioritisation of evidence, and
- specific adjustments for particular candidates.

Recommendation 4

More generally, in future, **robust exam-style evidence should be more habitually collected** within the course of study. This would embed contingency within the assessment

system for situations when terminal assessment is not possible. There are various ways by which this could be done, such as a return to modular examinations, online mock exams, adaptive tests or appropriately moderated teacher assessment.

It will be crucial here to ensure that a) this material is robust evidence of candidate ability, involving the capacity for meaningful standardisation across candidates and centres, and b) that over-assessment does not take place.

Conclusion

In conclusion, the TAGs process provided assessment outcomes (GCSE, AS and A level grades) to candidates in what was a difficult situation for both teachers and students. The analysis of TAG submissions proved useful to us as we were able to gain an understanding of the approaches centres took to collecting evidence and determining their students' TAGs. Overall, there were many commonalities between centres in terms of their approaches to selecting evidence and making judgements, with many using assessment similar to the kinds used during typical GCSE sessions. However, there was a large amount of variation both between and within centres; this inevitably raises various questions about comparability of standards between centres.

We hope our findings can feed into discussions about the shape of England's assessment system (and similar systems) both in the short and longer term. By illuminating some of the challenges and opportunities associated with teacher assessment, we hope to contribute to discussions about both assessment in emergency situations, and the future of assessment after the pandemic. This includes issues such as how candidate performance can best be assessed, recorded and compared, and the roles of external and internal assessment in that. We hope our findings about the quality of the data we saw, and our recommendations related to the importance of clear guidance on what should be expected in similar future situations, can help ensure that future processes can be appropriately evaluated.

References

Delaney, C. (2005). *The evaluation of university students' written work.* British Educational Research Association Annual Conference, Glamorgan.

DfE/Ofqual. (2021). *Analysis of consultation responses: How GCSE, AS and A level grades should be awarded in summer 2021*. Ofqual.

Gill, T. (2019). Methods used by teachers to predict final A Level grades for their students. *Research Matters: A Cambridge University Press & Assessment publication*, *28*, 33-42.

Holmes, S., Churchward, D., Howard, E., Keys, E., Leahy, F., Tonin, D., & Black, B. (2021). *Centre Judgements: Teaching Staff Interviews, Summer 2020. Research and Analysis* Ofqual.

Holt-White, E., & Cullinane, C. (2021). *A Levels and University Access 2021. Research Brief*. Sutton Trust.

JCQ. (2021). *JCQ Guidance on the determination of grades for A/AS Levels and GCSEs for Summer 2021: Processes to be adopted by exam centres and support available from awarding organisations*. Joint Council for Qualifications.

Johnson, M., & Coleman, T. (2021). *Teachers in the Pandemic: Practices, Equity, and Wellbeing*. Cambridge University Press & Assessment Research Report. https://www.cambridgeassessment.org.uk/Images/639354-teachers-in-the-pandemic-practices-equity-and-wellbeing.pdf

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education*, *28*(1), 91-105.

Lada-Wilicki, J. (2021). The results are in: teacher-assessed grades go ahead. *The Independent Schools Magazine*, 5. https://www.independentschoolsmagazine.co.uk/PDFs/march-2021.pdf

Lee, M. W., & Newton, P. (2021). *Research and Analysis: Systematic divergence between teacher and test-based assessment: literature review*. Ofqual.

Leech, T., & Chambers, L. (2022). How do judges in comparative judgement exercises make their judgements? *Research Matters: A Cambridge University Press & Assessment publication*, *33*, 31–47.

McVee, M. B., Dunsmore, K., & Gavelek, J. R. (2005). Schema theory revisited. *Review of educational research*, *75*(4), 531-566.

Ofqual. (2020). *Summer 2020 grades for GCSE, AS and A level, Extended Project Qualification and Advanced Extension Award in maths: Guidance for teachers, students, parents and carers*. Ofqual.

Ofqual. (2021a). *Guidance for schools, colleges and other exam centres on contingency arrangements for students entering GCSEs, AS and A levels, the Advanced Extension Award and Project qualifications in summer 2022. Published 11 November 2021.* https://www.gov.uk/government/publications/guidance-on-contingency-arrangements-for-gcses-as-and-a-levels-in-summer-2022/guidance-for-schools-colleges-and-other-exam-centres-on-contingency-arrangements-for-students-entering-gcses-as-and-a-levels-the-advanced-extension

Ofqual. (2021b). *Guidance: Information for heads of centre, heads of department and teachers on the submission of teacher assessed grades: summer 2021 (HTML)*. Ofqual. Retrieved 21st March 2022 from https://www.gov.uk/government/publications/submission-of-teacher-assessed-grades-summer-2021-info-for-teachers/information-for-heads-of-centre-heads-of-department-and-teachers-on-the-submission-of-teacher-assessed-grades-summer-2021-html

Roberts, N., & Danechi, S. (2021). *Coronavirus: GCSEs, A Levels and equivalents in 2021 and 2022*. Commons Library Research Briefing 26th July 2021. House of Commons Library.

Vitello, S., & Williamson, J. (2017). Internal versus external assessment in vocational qualifications: a commentary on the government's reforms in England. *London Review of Education*, *15*(3), 536-548.
Williamson, J. (2018). Characteristics, uses and rationales of mark-based and grade-based assessment. *Research Matters: A Cambridge Assessment Publication*, *26*, 15-21.