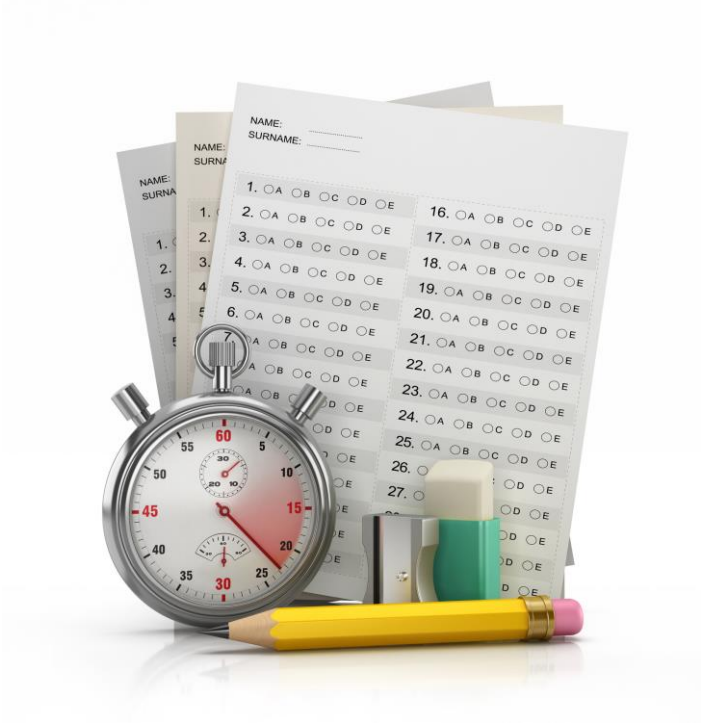
Examining exams

Are there credible alternatives to written examinations?

Tom Richmond and Eleanor Regan

April 2023



About the authors

Tom Richmond is the director of the EDSK think tank.

Tom has spent two decades in the world of education. He began his career teaching A-level Psychology at one of the country's leading state schools, having gained a BSc in Psychology from the University of Birmingham and an MSc in Child Development from the Institute of Education in London.

After three years in teaching, he moved into politics to work on policy development and research across the education, skills and welfare sector. This included roles at think tanks such as Policy Exchange and the Social Market Foundation, Pearson, G4S and working for an MP.

He subsequently spent two years as an advisor to ministers at the Department for Education, first under Michael Gove and then Nicky Morgan, where he helped to design and deliver new policies as well as improve existing ones. After leaving the Department for Education, he spent two years teaching at a Sixth Form College before moving back into education policy and research, first at the Reform think tank and then at Policy Exchange before deciding to launch EDSK.

He has also written extensively for publications such as the TES and Schools Week and has appeared on numerous media outlets, including the BBC News Channel, Sky News, BBC Radio 4, BBC Radio 5 Live, LBC and TalkRADIO.

Eleanor Regan is a researcher at the EDSK think tank.

She has co-authored several reports at EDSK, including major projects on the future of assessment in primary and secondary schools, the quality of apprenticeships and the debate over 'low value' Higher Education.

Before joining EDSK in 2021, she completed a BA in Geography at the University of Southampton, where she developed a strong interest in issues of inequality, particularly in relation to social class. In her spare time, she is an admin volunteer for a social enterprise that focuses on improving children's attainment in mathematics.

Acknowledgements

The authors would like to thank the three external peer reviewers of this paper who kindly provided their feedback and comments during the drafting process. Thanks also to those who commented on specific sections of the report or provided expert input for one or more of the recommendations.

The views expressed in the report are attributable to the authors and do not necessarily reflect EDSK's advisory board or supporters. Any errors remain the responsibility of the authors.

About EDSK

EDSK is an independent, impartial think tank.

Our mission is to design new and better ways of helping learners of all ages succeed, particularly those from less privileged backgrounds.

We aim to achieve this by conducting research on topics across the education and skills landscape and then producing evidence-based recommendations for how the system needs to change, both now and in the future.

Our research covers primary schools, secondary schools, colleges, apprenticeships and universities. Whichever topic or policy area we are investigating, our priority is always to produce better outcomes for learners while also ensuring that the greatest support is given to the most disadvantaged members of society.

We work closely with a range of individuals and organisations to ensure that our research reflects a wide range of evidence on any given issue. EDSK retains copyright and full editorial control over our research, irrespective of how it is funded.

EDSK is a not-for-profit think tank registered in England and Wales.
(Company Registration No. 11795717)

Find out more at www.edsk.org

Contents

Executive summary

Over 300 years since the first written exam was used in the English education system, this traditional form of assessment continues to divide opinion. To their supporters, written exams provide a rigorous test of students' knowledge and understanding that acts as a source of motivation as well as a sound basis for progression onto university or employment. Indeed, Prime Ministers, Education Secretaries, Schools Ministers and regulators have publicly stated that written exams are the 'best and fairest' way to measure pupils' attainment. Meanwhile, critics argue that written exams are narrow assessments that focus too much on memorisation and fail to provide students with the wide range of skills that they need for later life and work.

With a General Election looming, coupled with the collapse of the exam system in 2020 and 2021 due to the outbreak of COVID-19, debates over the future of exams have become increasingly vocal. As a result, this report set out to understand if the current dominance of written exams in our assessment landscape is justified and whether the following alternatives to exams could and should play a greater role in our high-stakes assessment system towards the end of secondary education - most notably at age 18:

- Coursework and controlled assessments
- Oral exams
- Portfolios
- Extended essays and projects
- Performance-based assessments

Developing and demonstrating a wide range of skills

A common criticism of written exams is that they focus too heavily on recalling knowledge, whereas other methods of assessment can emphasise other competencies. For example, the Extended Project Qualification (EPQ) – a voluntarily and independently-produced essay or project completed alongside A-levels – encourages students to investigate a topic of their choice, with the aim of developing their research, extended writing and presentation skills. Meanwhile oral assessments (such as the speaking components of language exams) give students the opportunity to demonstrate their knowledge in a more practical way while also seeking to improve their verbal communication skills.

Developing wider skills through different methods of assessment is not just a theoretical goal. This report identified several studies showing that a student's grade on the same course material may be different in an oral assessment or a portfolio (a collection of work, often used to assess subjects such as design and technology) compared to a written exam, indicating that these alternative assessments may be capturing different elements of performance. Using

'multi-model' assessment to get a broader view of a student's capabilities is common in technical education and apprenticeships but rarely features in academic settings.

Assessments that reflect 'real-world' settings

Written exams are normally completed in an artificial environment (such as a silent hall) that does not reflect real-world settings. In contrast, some alternative assessments allow students to acquire and demonstrate skills needed for employment and further study. For example, there is evidence showing that students who complete the EPQ may be better prepared for university degrees, while oral assessments promote the verbal communication skills that employers frequently claim are lacking among many school and college leavers.

The nature of some subjects means that assessments which closely resemble real-world settings are undoubtedly preferable to written exams. For example, the most appropriate way to assess a student's musical skills is through a live performance, while other artistic abilities such as drawing and painting are best captured through a portfolio of work. Although there is evidence to show that assessing creative subjects inevitably involves a greater degree of subjectivity (and thus less consistent grading) than a written exam, they remain the most credible way of capturing a student's attainment in these subjects.

Guarding against malpractice

When they were first examined in 1988, GCSEs often had a large coursework component. Just three years later, then Prime Minister John Major voiced concerns that standards were "at risk" with some students allegedly getting too much assistance from teachers or parents or even having their coursework written for them. To contain the risk of malpractice, there was a shift in the mid-2000s from coursework to 'controlled assessments' i.e. coursework completed under supervised conditions, with much tighter controls on its design, delivery and marking. Even so, a review in 2013 by the exam regulator Ofqual found that there was still "too many opportunities for plagiarism" and that, in some subjects, there was very little to distinguish between a controlled assessment and a written exam due to the tight controls. Consequently, Ofqual severely curtailed the use of 'non-exam assessment' (NEA) including coursework-style tasks. Many GCSEs and A-levels (e.g. history, geography, drama) have seen significant reductions in the contribution of NEA towards a student's final grade, and in some subjects (e.g. science) NEA has been eliminated altogether.

Despite these changes to GCSEs and A-levels, concerns over malpractice in a high-stakes assessment system persist in other forms of assessment such as the EPQ and the International Baccalaureate's (IB) compulsory 'Extended Essay', which are both completed without supervision. The development of new technology such as ChatGPT and other chatbots has exacerbated existing concerns regarding plagiarism as these tools can produce entire essays

and projects with minimal input (if any) from the student. Such is the inability of exam boards to identify malpractice related to chatbots, the IB recently announced that students were actually allowed to use such software to complete their Extended Essays. In contrast, the scope for malpractice in written and oral exams is greatly reduced by the controlled testing environment, thus making the grades awarded for these assessments more trustworthy.

The practicality of assessments in a high-stakes system

The inconsistencies in how coursework and controlled assessments were delivered in schools and colleges created numerous problems when seeking to award grades on a fair basis across the country, particularly when some students ended up receiving more advice and assistance than their peers (even within the same institution). Although controlled assessments sought to enforce more specific rules on the level of permitted help for students, teachers reported that there was still too much room for interpreting the rules differently and Ofqual found that in some cases "too much teacher input" continued. Extended essays and projects such as the EPQ continue to face the same challenges. For example, students must write their own research questions for their EPQ but teachers are allowed to provide feedback on them - meaning some students could be receiving more support than others, potentially giving them an unfair advantage. Written exams largely avoid these problems by ensuring that all students taking the exam receive the same questions in a standardised and strictly controlled environment, so the results should reflect a student's genuine attainment rather than being influenced by the amount of support that they received.

Written exams are also relatively cheap to deliver and mark, which is hugely beneficial when assessing tens (if not hundreds) of thousands of students over a short period. In comparison, coursework and controlled assessments were both very time consuming – often taking several months to complete – and used up a large portion of the curriculum time for each subject. Moreover, they increased the workload of teachers who had to supervise and mark the tasks. Other methods of assessments also need a considerable amount of time to ensure that they produce a credible measure of student attainment. Research has shown that the consistency of marks awarded for portfolio assessments improves when multiple assessors mark each collection of work produced by a student (as is commonly done in smaller subject such as art and design), but this intensive approach would quickly become unfeasible at a larger scale.

The challenges caused by asking teachers to award grades

Students often appear to perform better in assessments such as coursework and controlled assessments that are graded by their teacher rather than external examiners, yet research by Ofqual found that this was not necessarily a "fair representation" of a student's attainment. Asking teachers to award grades for these assessments also made it hard to differentiate between students because of a 'bunching' towards the top end of the available marks. This

was perhaps unsurprising as teachers reported being in a "difficult, sometimes stressful" position when marking their students' work as they knew that "their own performance and that of the school" would be judged by the results. The EPQ, which is marked by teachers, has seen a similar bunching of marks, with 45 per cent of candidates being awarded an A or A* in 2019 (rising to 55 per cent during the pandemic) compared to 25 per cent across all A-levels.

Written exams typically produce grades that expose different levels of attainment within a cohort of students because they are externally mark and designed to test students' knowledge of externally set content. These normal safeguards disappeared during the pandemic, leaving teachers with the unenviable task of determining their own students' grades. The proportion of A and A* grades awarded across all A-level subjects subsequently leapt from 25 per cent in 2019 to 44 per cent in 2021, with sharp rises in top grades also visible in GCSEs. Teachers were put under immense pressure during this period and reported frequently working late into the night to manage their substantial workload due to this enforced experiment with teacher-assessed grades. A survey by Ofqual in 2021 found that less than 40 per cent of the public had confidence in A-level and GCSE grades awarded during the pandemic, further emphasising the risks created by a grading system that does not produce trustworthy outcomes. When coupled with the findings of numerous inquiries and studies conducted well before the pandemic, the research evidence clearly demonstrates why asking teachers to award grades to their students should be avoided within a high-stakes assessment system.

Widening of disparities between groups

Numerous research studies have found that when teachers are asked to award grades to their students, they can be influenced by their existing knowledge of that student. For instance, a teacher may inadvertently award a piece of work a higher grade than it deserves because the student in question is generally a high achiever. Teachers can also be influenced by a pre-conceived (and often subconscious) idea of how well a student may perform based on demographic factors, with multiple studies showing that grades awarded by teachers can be lower for children from less well-off families and those with special educational needs compared to other children of the same ability level. When teachers were responsible for awarding grades during the pandemic, concerns over widening disparities between students were again evident as some existing performance gaps between students from different demographic groups increased – particularly for black students and those from lower socio-economic backgrounds.

This so-called 'bias' is not intentional, nor is it unique to teachers, as even external assessors can be biased in their judgements. For example, one study found that during musical performances an assessor's mark can be influenced by the gender and ethnicity of the performer as well as the experiences of the assessor themselves (e.g. how familiar they are

with the piece being performed). In contrast, written exams reduce the opportunity for bias to occur as they are marked anonymously as well as externally. Other methods of assessments can also limit the opportunities for bias by using anonymous marking (as is done for marking A-level and GCSE music performances as well as the IB Extended Essay).

Different approaches to marking

Written exams are generally judged in a consistent way as assessors follow the same mark scheme that sets out the knowledge and skills required from candidates. This approach makes it more likely that if a student were to take the exam again, they would achieve an identical or very similar grade. Some alternative methods of assessment struggle to achieve the same consistency, as demonstrated by the pandemic-era experiment with teacher-assessed grades. Research has shown that consistent grading can also be difficult to achieve when the assessor is asked to make a more subjective judgement (e.g. art portfolios or drama performances) because even with a mark scheme, assessors may value different skills, traits and styles in creative outputs.

Several studies have described ways to improve consistency between assessors. One of the most important findings is that there is greater consistency between teachers when they are asked to rank students in order of how well they performed rather than awarding specific grades. Another way to improve consistency is by asking the assessor to make their judgement in a different way. Studies have found that asking assessors to award a single holistic score to a piece of work in the absence of any formal marking criteria can often produce more consistent grades than asking assessors to judge the same work using a detailed and prescriptive mark scheme.

Conclusion

There are good reasons why written exams have come to dominate the assessment system in schools and colleges. They are a relatively low-cost, standardised and impartial way to assess students' knowledge and understanding, with a much lower probability of being affected by malpractice or inconsistent grading than other methods of assessment. The controlled setting in which exams normally take place also means that students, parents, universities, employers and the government can have confidence that the awarded grades are a genuine reflection of a student's attainment. Any reforms that may result in greater inaccuracies or inconsistencies in grading would be detrimental to students as well as taxpayers who have every right to expect a publicly funded system to deliver fair judgements on students. Nevertheless, every method of assessment involves trade-offs and written exams are no exception, particularly their limited value in building many skills that are useful beyond the confines of an exam hall.

Regardless of the imperfections of written exams, the problems faced by many alternative forms of assessment are hard to ignore. The advent of ChatGPT is a significant threat to the integrity of formal assessments in this country and elsewhere. Plagiarism has always been a risk to some extent, especially for coursework-style tasks, but establishing for certain whether a student produced the work that they submitted has now become a virtually impossible task for teachers, leaders and exam boards. Consequently, it would be unwise to increase the proportion of coursework or similar assessments into our high-stakes system because there is no realistic prospect of preventing widespread malpractice. What's more, the burden that would be placed on teachers by switching from external exams to more internal assessments should not be underestimated given what teachers have consistently reported in the past.

This report concludes that both supporters and critics of written exams make valid arguments regarding the benefits and drawbacks of this enduring form of assessment. As a result, the following recommendations seek to build on the most commendable attributes of written exams while also drawing on the benefits of other types of assessment that can withstand the demands of our high-stakes assessment system. If, as this report proposes, a government – either current or future – is willing to invest more in schools and colleges to ensure that every institution can offer a wider range of courses and assessments, our assessment system will be placed on a stronger foundation for many years to come.

Recommendations

- **RECOMMENDATION 1**: To maintain the credibility of the high-stakes assessment system in the final years of secondary education, written examinations should continue to be the main method of assessing students' knowledge and understanding. In contrast, placing a greater emphasis on coursework and other forms of 'teacher assessment' would increase teachers' workload and lead to less reliable grades that may be biased against students from disadvantaged backgrounds.
- **RECOMMENDATION 2**: To broaden the curriculum and develop a wider range of skills than those promoted by written exams, students aged 16-19 taking classroom-based courses should be required to take one additional subject in Year 12 (equivalent to an AS level) that will be examined entirely through an oral assessment.
- **RECOMMENDATION 3**: To ensure that students taking classroom-based subjects can develop their research and extended writing skills beyond an exam setting, the Extended Project Qualification (EPQ) should be made compulsory. In future, the EPQ will be used as a low-stakes skills development programme and will therefore be ungraded.
- **RECOMMENDATION 4**: To give schools and colleges the resources they need to expand their 16-19 curriculum to include an additional subject and the EPQ, the 'base rate' of per-student funding (currently £4,642) should be increased by approximately £200 a year to reach £6,000 by 2030.

1. Introduction

> *"One had to cram all this stuff into one's mind for the examinations, whether one liked it or not. This coercion had such a deterring effect on me that, after I had passed the final examination, I found the consideration of any scientific problems distasteful to me for an entire year."* [1]
>
> **- Albert Einstein**

Despite having some famous critics over the years, it is no exaggeration to say that written examinations dominate assessment in English schools. With few exceptions, A-level and GCSE qualifications are built around pen-and-paper assessments conducted in silence and lasting up to three hours. Other forms of assessment such as coursework have fallen out of favour due to their supposed lack of rigour, the workload burden they place on teachers and their perceived failure to produce reliable and consistent grades - leaving written exams in a seemingly unassailable position. The rejection of most other assessment methods is due in no small part to the belief – repeatedly expressed by Prime Ministers, Education Secretaries, Schools Ministers and the exams regulator Ofqual – that written exams are the 'best and fairest' way to measure pupils' attainment.[2]

Considering their prominence in the modern era, it is perhaps surprising that written exams are a relatively recent invention by historical standards. The first known written assessments are generally agreed to be the Imperial Chinese examinations, starting in the year AD 606.[3] These offered entry to a well-paid and highly-esteemed career in the Civil Service and were entirely meritocratic, with no restrictions on who could enter or how many times the exams could be sat.[4] The most demanding component (the 'jinshi' degree) tested examinee's knowledge of the Confucian classics, history, proficiency in compiling official documents, inscriptions, discursive treatises, memorials, and poems and rhapsodies.[5] Meanwhile, examinations in Europe in the early Middle Ages generally involved little more than learning and reciting religious texts along with simple questions-and-answers.[6] As the first European universities were established in Paris, Bologna and elsewhere, examinations began to incorporate other methods such as delivering lectures and taking part in public 'disputations' – a precursor to the modern viva.[7] The disputation has been described as the 'high mark' of medieval education. It normally involved one of the 'masters' posing a theorem or problem, after which one student (who was being examined) would defend the idea while others (either masters or students) could oppose them. Such was the intensity, depth and scrutiny generated by a disputation, they could last an entire day.[8]

By the 1600s, all examinations at Cambridge and Oxford Universities were still oral assessments (in Latin) where candidates were challenged in public by senior university staff. However, by 1700 some students were faced with written examinations that were intended to 'weigh and compare the suitable merits of the young men with suitable deliberation'.[9] Trinity College, Cambridge, is thought to have created the first written examination in Europe in 1702,[10] marking the start of a concerted movement away from oral assessments. By 1722, questions at Cambridge were being dictated to students for written answers. From 1790, some exam papers were printed and students could take the papers away to complete them. By 1828, all papers were printed and examiners had only limited opportunity to conduct oral assessments. Oxford University followed suit by introducing written exams in the 1820s, although their institutional focus on theology and classics (as opposed to Cambridge's focus on mathematics and science) continued to draw on oral assessments to some extent.[11]

There were three driving forces behind this shift towards written examinations. First, as Oxford and Cambridge began to expand, there was less space and time available for disputations and other oral assessments. Second, written tests were a private and silent way of examining a larger number of students, in contrast to the public and verbal disputations that preceded them. Third, the shift reflected the desire to rank students rather than simply place them into broad categories of performance.[12] The landmark report on school examinations by former MP Arthur Herbert Dyke Acland in 1911 (the 'Acland Report') underscored the significance of these changes. Acland stated that "examinations began as a method of testing the efficiency of a candidate for the practice of some profession or for admission to some learned society", but in the 18th century "examinations took on a new function, namely, the distinguishing between candidates for academic distinction according to their different degrees of intellectual merit".[13]

This desire to rank students soon became more consequential, as many scholarships and other university distinctions and opportunities were available only to the top-performing students. In this competitive environment, low-achieving university students could hide their shame by taking a 'pass', in which case they were not ranked and their names were not made public. To illustrate what was available to those who embraced the new preoccupation with ranking, participants in the revered 'Mathematical Tripos' at Cambridge University were faced with 16 papers spread over eight days, giving them 44.5 hours to answer 211 questions with a total possible mark of 17,000.[14] Until 1910, Tripos participants were ranked according to achievement – the best being honoured with the title of 'Senior Wrangler', and the lowest receiving a title that was intended to resemble a man-sized attribute: a 'Wooden Spoon'.[15]

Throughout the 19th century written exams became commonplace across a variety of careers and occupations, taking a central role in medicine, pharmacy, veterinary science, chemistry, engineering, the army and the navy and the legal profession.[16] In 1870, written examinations

were introduced into the Civil Service to promote free competition for places as "a remedy for the system of patronage and jobbery under which nearly all the posts in the service of the State had previously been allotted"[17] (similar to the aim of written exams in Imperial China). It was around this time that Oxford and Cambridge both established examinations for young people under the age of 18, with the first sat in 1858.[18] In 1870, the newly-established Headmasters' Conference wanted the Government to create a system of 'leaving examinations' for school pupils, but in the end they invited Oxford and Cambridge to join forces – leading to the Oxford and Cambridge Schools Examination Board in 1873.[19] In the following decades, numerous universities such as Leeds, London, Sheffield, Manchester and Liverpool began to use their matriculation (entry) examinations as school-leaving exams as well.[20]

By the time the Acland Report on school examinations was published in 1911, the situation had become rather chaotic. Many universities as well as professional and commercial bodies were producing and conducting their own examinations for pupils of different ages to determine their suitability for certain courses and occupations. These organisations did not ordinarily recognise each other's examinations as 'equivalent' and often refused to accept them as sufficient preparation for their own institution. The Acland Report had "no hesitation …in stating our conviction that external examinations are not only necessary but desirable in secondary schools",[21] although it demanded that the "existing multiplicity of external examinations (including those of universities, and professional and other bodies) …should be reduced by concerted action."[22] As a result, the 'Secondary School Certificate Examination' was rolled out by the government in 1917,[23] which mainly consisted of written tests in various subjects but also included practical and oral elements where appropriate.[24] Those who chose to stay on until age 18/19 took the 'Secondary School Higher Certificate Examination', which was "not only for those who are proceeding from school to the university, but also for those who are intending to follow a professional or commercial career after leaving school."[25]

These assessments remained in place until 1951 when the General Certificate of Education (GCE) was introduced. GCE exams were available at Ordinary Level (O-level; equivalent to the Secondary School Certificate) and Advanced Level (A-level; equivalent to the Secondary School Higher Certificate). Unlike the Secondary School Certificate Examination that required pupils to pass a group of subjects, the GCE system allowed them to sit and pass individual subjects.[26] The GCE was explicitly aimed at high-ability pupils in private and grammar schools, meaning that most pupils at secondary moderns ended up leaving school without any recognised qualifications. In response, the Certificate of Secondary Education (CSE) was introduced in 1965 to provide a set of qualifications that were distinct from O-levels by covering both academic and vocational subjects, incorporating teacher-assessed components alongside written examinations and having exam questions that were typically shorter and more structured than O-Level papers.[27] The CSE struggled to gain credibility among policymakers, parents and employers[28] and the raising of the school leaving age to 16 in 1972[29]

ensured that O-levels became the main qualification within secondary education. O-levels and CSEs were subsequently swept away in 1986 by the creation of the General Certificate of Secondary Education (GCSE) for 16-year-olds, who could then leave school or progress onto A-levels or equivalent courses.

Fast forward to the present day and, while written exams are certainly not universally loved, they have become central to the operation of two prominent elements of education in England: the high-stakes accountability system for schools and the competitive application process for university entrance. As with written exams themselves, both elements have their critics. Even so, the school accountability system and competitive university applications are likely to remain in place for the foreseeable future, which is why this report positions its research and analysis firmly within a high-stakes environment. Consequently, the underlying focus of this report will be on summative (final) national assessments in the later years of secondary education – most prominently at age 18. Given the high stakes attached to the outcome of these assessments, there would need to be a compelling case for switching away from written exams to other forms of assessment because ultimately it is the learner who loses out if their formal assessments produce less accurate or trustworthy judgements on their ability. What's more, employers and universities need to have confidence in the results of national assessments that feed into their application processes to ensure that they can select the most suitable (although not necessarily the highest performing) candidates.

This report opens with a consideration of the benefits and drawbacks of written examinations. Following this, a wide range of alternative assessment methods will be explored to understand where they may (or may not) be able to add value in a high-stakes system. To determine each assessment method's potential value, a detailed analysis will be conducted of the most important messages emanating from academic studies, research reports and other independent sources of evidence. Throughout this analysis, the respective merits of each alternative to written exams will be investigated in relation to the following five attributes:

1. **VALIDITY**: the extent to which the assessment method measures what it claims or intends to measure;

2. **RELIABILITY**: the extent to which the level of performance recorded by an assessment is consistent from one use of the test to the next (this is normally reported on a scale of 0 (no consistency) to 1 (perfect consistency));

3. **REAL-WORLD APPLICABILITY**: whether the assessment method requires students to demonstrate the same competencies, knowledge, skills and aptitudes that they would need to apply in a real-world or professional setting;

4. **PRACTICALITY**: the extent to which the assessment method is viable on a large - potentially national - scale when judging the performance of thousands of students (e.g. cost of delivery; workload burdens on teachers).

5. **CREDIBILITY**: whether the assessment method produces a judgement on a student's performance that can be trusted by others (e.g. employers and universities).

After discussing the various alternatives to written exams, this report will put forward a set of proposals for how summative assessment in secondary education could be reformed within a high-stakes environment. It is therefore hoped that this report makes a valuable contribution to the debate over the future of assessment in England.

2. The benefits and drawbacks of written examinations

As will be described throughout this report, every method of assessment in mainstream education involves some form of trade-offs, and written exams are no exception. Consequently, this chapter will briefly explore the main reasons why written examinations are cherished by many educators before describing some common criticisms levelled at them.

Benefits of written examinations

- **A direct test of knowledge**: the most frequently cited strength of written exams is that they offer a direct assessment of a student's knowledge on any subject or topic as well as their ability to express said knowledge in a comprehensible manner. As the Acland Report in 1911 put it, written exams make students "work up to time by requiring him to reach a stated degree of knowledge by a fixed date" and "incite him to get his knowledge into reproducible form and to lessen the risk of vagueness".[30]
- **Objectivity**: the widespread use of anonymous marking for formal written examinations ensures that they avoid the potential for subjective judgements when the assessor knows the student being assessed. This perceived objectivity is one of the main reasons why written exams are used for academic selection (e.g. university entrance) and have been adopted by many professions and occupations to remove the influence of patronage and personal connections when allocating jobs and opportunities.
- **Standardisation**: although written exams vary in their content and structure, they are typically standardised assessments i.e. all those taking the assessment receive the same questions under the same conditions (e.g. a silent hall) and are theoretically judged by markers in a consistent manner. This tends to improve the validity of the test as a direct measurement of performance and it also improves the reliability (consistency) of scoring, giving governments, pupils and parents more confidence that students would be likely to achieve an identical or very similar mark or grade if they sat the test again.
- **Supporting comparisons**: written exams normally produce a raw score that can be used to place students in rank order as well as assign them a numerical or letter grade. The Acland Report recognised that this allows a student to measure their attainment "(i) by the standard required by outside examiners, (ii) by comparison with the attainments of his fellow pupils, and (iii) by comparison with the attainments of his contemporaries in other schools."[31] The ranking of cohorts of students also allows universities and employers to make inferences about a student's current and future capabilities. Meanwhile, the scores for entire cohorts allow governments to monitor results within institutions, between institutions and across a whole country in a manner that other assessment methods

struggle to match – particularly the ability to monitor standards in schools and colleges over time as part of the accountability system.

- **Motivating students**: because written exams provide a direct assessment of a student's knowledge and understanding, they can provide a source of motivation by focusing students' efforts on a specific goal. This should theoretically provide an incentive for students to study harder, while the exam setting itself may also test a student's ability to work under pressure.
- **Clearly articulated standards and content**: unlike some other forms of assessment that allow candidates to choose their own topic or material, written exams are normally based on an agreed curriculum or programme that sets out the depth of knowledge and understanding that students must reach depending on their age and the subject in question. This articulation of standards and content tends to support the validity and reliability of written exams as a measure of how much learning has taken place (which is valuable for teachers) as well as supporting fairer comparisons in terms of what students have achieved.
- **Relatively low-cost**: the lower cost of written exams gives them a considerable advantage over other forms of assessment throughout primary, secondary and tertiary education. When assessing hundreds, if not hundreds of thousands of students in a short timeframe, many other assessment methods would not be able to offer a reasonable level of validity and reliability without requiring a prohibitively high level of financial investment.

Drawbacks of written examinations

- **Focus on knowledge recall and rote learning**: the Acland Report was concerned that written exams risked "favouring a somewhat passive type of mind" by "rewarding evanescent forms of knowledge." In addition, a focus on knowledge recall encourages students to aim for "absorbing information imparted to him by the teacher", which may result in "setting a premium on the power of merely reproducing other people's ideas".[32]
- **Artificial testing conditions**: few occupations and careers require individuals to demonstrate their knowledge in silent written assessments conducted over a few hours, even if they may need to pass exams as part of their training. In other words, written exams do not reflect the real world because they are completed in an artificial environment that does not relate to how learners will use their knowledge and understanding in Higher Education or a professional context.
- **Missing wider skills and aptitudes**: a common complaint from employers is that the examination system does not give students the skills they need to thrive in the workplace. The recent *Employer Skills Survey* by the DfE found that a 'lack of the required soft / personal skills or competencies (e.g. problem solving, communication or [teamwork])'

was one of the most frequently absent skills among education leavers.[33] Written exams can struggle to credibly assess these wider skills, unlike the broader emphasis on *Knowledge, Skills and Behaviours* that form the basis of apprenticeships in England.[34]

- **Demotivating effects**: while some students may be motivated by written exams, others claim they are stressful and even detrimental to their mental health.[35] Students who are adversely affected by exams may not perform at their full potential – thereby reducing the validity of the exam as a measure of performance. That said, it is hard to calculate how many students may be affected in this manner (or in which subjects / disciplines), and both anxiety and stress could potentially be reduced by the actions of students and teachers in many cases.
- **Teaching to the test**: teachers 'drilling' their pupils in the subject material on which they will be assessed, along with devoting a large proportion of time to test preparation, exam techniques and even question-spotting, can reduce the validity of written exams. For example, as far back as 2008 a parliamentary committee received "substantial evidence that teaching to the test …is widespread"[36] while Ofsted, the school inspectorate, has found that the curriculum can be narrowed to the point where teachers and students are only willing to put effort into learning content that is likely to appear in an exam.[37]

It is clear, then, that written exams in schools and colleges offer many benefits for students, government ministers, parents, universities and employers, particularly their standardised nature and objectivity. Even so, they are no panacea from an assessment perspective. The most obvious trade-off with written exams is that they tend to prioritise standardisation, impartiality and improving reliability at the expense of reflecting the 'real world' and supporting wider skills development in a credible manner. What's more, research from the exam regulator Ofqual has shown that written exams can still produce inconsistent grading in some subjects,[38] further emphasising how written exams – like every other assessment method – have their imperfections. Nevertheless, within the context of a high-stakes assessment system, written exams have set a relatively high bar for the trustworthiness of the final grades awarded to students. The question for the various alternatives to written exams discussed in the remainder of this report is therefore as follows: to what extent could they offer as much assurance as the grades achieved in written exams while addressing some of the weaknesses associated with written exams?

3. Coursework and controlled assessments

Calls for an increased focus on coursework have come from different parts of the political spectrum. Former Conservative Education Secretary Lord Baker has called for GCSEs to be replaced by a mixture of exams and coursework,[39] while former Labour Education Secretary David Blunkett has expressed his desire to measure "continuous learning" rather than relying on end-point exams.[40] A YouGov poll had previously found that 64 per cent of the public were in favour of dividing marks between coursework and final exams, with only 28 per cent in favour of marks being based solely on final exams.[41] However, critics have argued that "replacing exams with coursework just creates a new kind of unfairness"[42] and that previously teachers and pupils found coursework "stressful and burdensome".[43] To investigate these respective arguments, this chapter will explore the history of coursework in England in recent decades.

The early challenges facing GCSE coursework

In 1988 when GCSEs were initially introduced, assessment in almost all subjects consisted of a combination of written exams and coursework. Coursework was used to assess skills that were difficult to evaluate accurately through written exams such as carrying out practical experiments, creative performances and writing extended essays.[44] As a result, coursework aimed to give students the opportunity to work at their own pace and take responsibility for their learning as well as studying a topic in depth. Teachers were also able to set specific tasks that suited the level and interest of individual students.[45] Some coursework was done outside of school hours, while some was done under supervision in school. Exam boards had moderation processes in place to check teachers were applying the mark scheme for coursework tasks appropriately and consistently.[46]

Even so, the use of coursework was not universally supported. Wariness about its use arose when one English GCSE was assessed entirely through coursework and became so popular that most 16-year-olds were taking it.[47] Just three years after the first GCSEs were introduced, then Prime Minister John Major gave a speech acknowledging concerns regarding the reliance on coursework:

> *"…[there is] suspicion that standards are at risk. It is clear that there is now far too much coursework, project work and teacher assessment in GCSE. The remedy surely lies in getting GCSE back to being an externally assessed exam which is predominantly written."*[48]

He went on to propose that a maximum of 20 per cent of marks should be obtainable from coursework. Following the speech, lower limits were indeed introduced although they were generally not as low as 20 per cent except for subjects like mathematics and religious studies, with subjects like English and art maintaining a higher weighting at 40 per cent and 60 per cent respectively.[49]

By the early 2000s, further questions over the credibility of coursework were being voiced in terms of marking reliability, the authenticity of pupils' work, teaching practices and the impact of coursework on teaching time.[50] Media reports at the time included claims that some teachers were "routinely writing the coursework" for their GCSE pupils or getting pupils to copy their work (sometimes under pressure from senior leaders at their school).[51] As a result, the Qualifications and Curriculum Authority (QCA) carried out a review of GCSE coursework arrangements in 2005. They concluded that the benefits of coursework outweighed the drawbacks but identified some ways coursework could be improved including better guidance for teachers on setting coursework tasks, clearer guidelines on the limits of 'permitted help' and advice from teachers and parents, and more checks by exam boards on schools' internal standardisation of marks. [52]

The move towards controlled assessment

In 2006, the QCA announced that while subjects such as art and PE would continue using coursework, there would be a movement towards new "controlled assessments" for subjects such as English literature, geography and history:[53]

> *"Controlled assessments will be taken under supervised conditions and will either be set by the awarding body and marked by teachers or set by teachers and marked by the awarding body. Controlled assessments may involve different parameters from those used in traditional writing examinations. They may, for example, allow access to sources such as the internet but under supervision."*

Following an independent report in 2007 by Dr Ian Colwill (commissioned by the QCA), more changes were introduced for the new GCSEs to be first taught in 2009. These included predetermined levels of control / supervision (limited, medium and high) applied at three stages: task setting, task taking and task marking. At each stage, the QCA aimed "to set the level of control as high as possible, to ensure the authenticity of students' work, while also attempting to make the assessments manageable in practical terms."[54]

In 2011 the newly established exam regulator Ofqual (which took over the QCA's regulatory functions) commissioned a survey to review teachers' experiences of controlled assessment.[55]

The survey found that, in general, the principles of controlled assessment were "well received" and that respondents were broadly supportive of the idea. Most teachers felt controlled assessment guarded against malpractice, provided a fair assessment of performance and assessed a broad range of skills.[56] However, there were several limitations including its implementation being more problematic in some subjects, concerns about the impact on teaching and learning time, and guidance for schools being ambiguous. Although teachers were largely positive about controlled assessment, it was clear that there were still "deep-seated concerns".[57]

In light of these limitations and a marking debacle in GCSE English in 2012,[58] Ofqual launched a comprehensive review of controlled assessments in 2013 that coincided with the Coalition Government's wider reforms to GCSE and A-levels.[59] The review uncovered five major issues:[60]

1. *"Many GCSE subjects include subject-specific elements that cannot be effectively assessed through written exams, but in reality the need for higher levels of control means that this is not always what is assessed by the current controlled assessment."*

This was a particular problem in geography as many schools were completing fieldwork exercises in a single day. In addition, Ofqual found that "the freedom to choose, plan, research and write up their work allowed too many opportunities for plagiarism, writing frames and too much teacher input".[61]

2. *"In some respects, controlled assessment has proved to be a better form of internal assessment than coursework, but the tighter controls have led to greater inconsistency in the way controls are implemented and the way work is carried out."*

Ofqual's survey found that many teachers felt there was scope for schools to "interpret the guidance differently", meaning that students were still not on a "level playing field".[62]

3. *"Controlled assessment presents practical difficulties for schools to manage and has had a negative impact on a number of aspects of teaching and learning."*

Ofqual found that controlled assessment was "not seen as an assessment that can be relied on to produce a fair representation of what students can do" and was instead "often treated as a hurdle that must be cleared".[63] 73 per cent of teachers felt that controlled assessment was not encouraging breadth and depth in teaching[64] and teachers also told Ofqual that it generated significant burdens, particularly for subjects with written controlled assessments or controlled assessment worth 60 per cent of the marks.[65] On a related note, teachers felt that they were being put "in a difficult, sometimes stressful, position when marking controlled assessment

work" as they knew that "their own performance and that of the school will be judged" by the results.[66]

4. *"In some subjects there is very little to distinguish between the controlled assessment task and the exam."*

The introduction of additional controls relative to coursework brought controlled assessment "much closer to a written exam",[67] with students completing their work in a supervised classroom environment not unlike the exam hall. In art and design, controlled assessments were essentially being used as "an opportunity to practice for the exam"[68] rather than demonstrating different skills.

5. *"Controlled assessment does not generally differentiate well between students of different abilities."*

As with coursework before it, a "bunching" of marks towards the higher end of available marks was found with controlled assessments. Data from GCSE geography and French showed that students tended to score more highly in controlled assessments than in the exam.[69] 72 per cent of respondents to Ofqual's call for evidence felt controlled assessment was not stretching and challenging students, while also not being appropriate for the less able.[70]

The move towards 'non-exam assessment'

The difficulties with controlled assessment led Ofqual to propose a set of principles for what they now called "non-exam assessment" (NEA) – which, as the name suggests, refers to any assessment that is not sat under standard exam conditions (e.g. taken simultaneously by all candidates) and therefore includes coursework, portfolios and performances. Ofqual's principles for NEA were as follows:[71]

- Non-exam assessment should only be used when it is the only valid way to assess essential elements of the subject;
- Non-exam assessment must strike a balance between valid assessment of essential knowledge and skills, sound assessment practice and manageability;
- Any non-exam assessment arrangements should be designed to fit the requirements of the particular subject including the relative weighting of written exams and other components assigned to it; and
- Non-exam assessment should be designed so that the qualification is not easily distorted by external pressures from the wider system.

Ofqual also added two caveats to these principles. Firstly, where NEA was to be used, Ofqual would specify the weighting, assessment objectives and focus of the assessment to ensure comparability between exam boards. For example, GCSE music has a 60 per cent weighting that must be equally split between composition and performance. Secondly, where NEA was the most valid form of assessment but its accuracy could potentially compromise the fairness of the exam result, then the outcome may be separated from the exam itself.[72]

Ofqual's principles led to a reduction in NEA in GCSEs. English, maths, the sciences, geography and history no longer have any NEA, while other subjects such as physical education and design and technology had the weighting of NEA reduced by 10 to 20 percentage points. Language GCSEs such as French, German and Spanish had their weighting reduced from 60 per cent to 25 per cent. In contrast, art and design GCSE continued to use 100 per cent NEA to reflect the nature of the subject and the validity of this type of assessment.[73]

In addition to the overall reduction in the amount of NEA within GCSEs, Ofqual made additional changes to two specific subjects in an attempt to improve validity. In GCSE English language, the marks from the spoken language aspect of assessment would not count toward pupil's overall grade, but would be reported separately from the written exam on a three-point scale.[74] For GCSE science, pupils would only be assessed by exam but at least 15 per cent of the total marks available were related to demonstrating understanding of scientific experimentation, and both pupils and schools have to keep records of their practical work.[75]

When Ofqual's NEA principles were applied to A-levels, fewer changes were required. Many subjects were left unchanged, with art and design remaining at 100 per cent NEA while subjects such as psychology and history continue to have none. Some subjects like English literature, English language and drama saw small reductions in their NEA weighting of up to 20 per cent[76] while geography saw the introduction of 20 per cent NEA in the form of fieldwork to reflect its importance within the subject.[77] Although for science subjects the NEA was reduced from 20 per cent to zero, students would instead carry out practical work and receive a 'pass' or 'fail' grade for this work that is separate to their written exam grade.

Ofqual's decision to remove practical work from a student's final science grade was a direct response to the "increasing number of allegations of malpractice in the conduct of these assessments" as well as concerns around the marking of practical work that had resulted in students doing "much better in them than in their exams". Moreover, the marks awarded for practical work had failed to "discriminate well between students" – a longstanding problem with coursework – while exam boards had no "verifiable evidence" of students' practical skills and could therefore not moderate teachers' marking effectively.[78] Although practical work would no longer contribute towards a student's final grade, their written exam could include questions related to the context of practical activities – as seen in GCSE science exams.[79]

This desire to require practical work but not include a student's performance in this work within their final grade is an example of how assessment can potentially be used to improve young people's skills without interfering with the validity or reliability of the assessment process. A recent study by Ofqual compared the 'hands-on' practical skills of university students, some of whom had completed A-level science subjects when practical skills were still part of a student's final grade and some of whom had completed the reformed A-levels without graded practical experiments. No statistical difference was found between the practical skills demonstrated by the two groups of students in chemistry or physics, while the students who took the reformed A-levels actually outperformed their peers in biology.[80] Although this research only included those students who went on to study science at university,[81] it provides some cause for optimism that there has not been a decline in the skills of A-level science students since practical experiments were removed from their overall grade.

Although the recent decisions by Ofqual regarding the level of permissible NEA have protected some element of 'coursework' in several subjects, the overall pattern of significant reductions in coursework is a clear indication of the demise in its credibility as an assessment tool. Despite numerous attempts by policymakers and regulators over the last two decades to ensure that the grades awarded for coursework and similar activities were valid and reliable, the evidence strongly suggests that such assessments struggle to withstand the pressure of a high-stakes system. Countless (and well-intentioned) rules and regulations were brought in over many years to manage how coursework and controlled assessments were delivered, but this seemed to cause more confusion and complications for teachers, not less, without offering much protection against malpractice. As a result, Ofqual's current principles for the use of NEA appear sensible given that any coursework-style activity which does not meet these principles has repeatedly failed to provide the necessary reassurance to pupils, parents and policymakers over the trustworthiness of the final grades awarded.

As if the previous problems with malpractice were not enough to restrict the use of coursework in the present day, new artificial intelligence (AI) tools such as ChatGPT and other chatbots can theoretically produce entire essays and projects with minimal input (if any) from a student. The credibility of the final grades awarded for any work produced independently by students is now undeniably threatened by the inability of assessors to confirm whether a student was in fact the sole author of a piece of work. Although this was a potential issue before the rise of chatbots (e.g. a student purchasing an essay from the internet or receiving assistance from a family member), the situation has become demonstrably worse in a very short timeframe.

Some schools have already opted to scrap homework amid fears that chatbots could be used to cheat on these tasks as well as formal exams,[82] while some leading universities around the world have elected to ban the use of ChatGPT altogether.[83] Ofqual's chief regulator Dr Jo

Saxton also recently stated that if she was a school leader she would make students do coursework under exam conditions to reduce the likelihood of cheating (reviving talk of controlled assessments), adding that ChatGPT "reinforces the importance" of exams that "have stood the test of time so well".[84] Although new guidance has been produced by exam boards to address concerns related to the use of AI tools, there is no doubt that malpractice would rapidly become widespread if a greater proportion of a student's grade was allocated to tasks that are completed outside of a controlled exam setting.

4. Teacher assessment

Many of the problems discussed in relation to coursework and controlled assessment in the previous chapter stem from the fact that they are typically marked by a pupil's own teacher, albeit with external moderation processes used in certain contexts. Perhaps the most extreme example of 'teacher assessment' came during the COVID-19 pandemic, when teachers were asked on two separate occasions to determine their pupils' grades with minimal oversight from an external body. This chapter will explore the impact of this natural experiment with teachers awarding grades to their own pupils.

Teacher-assessed grades during the pandemic

In March 2020, the Government made the "difficult" decision to cancel all exams due to take place in schools and colleges in England that summer.[85] Ofqual was subsequently tasked with developing a process to provide a calculated grade for each student reflecting their performance "as fairly as possible".[86] Teachers would be required "to submit judgements about the grade that they believe students would have received if exams had gone ahead",[87] as well as ranking their pupils within each grade for each subject. This information would be combined by exam boards with other relevant data including pupils' prior attainment, and the calculated grade would then be produced. A similar approach was proposed for vocational and technical qualifications, although qualifications that were used to signal occupational competencies had to be adapted or delayed.

A-level results day in August 2020 saw a significant backlash against Ofqual's 'standardisation' model, as 36 per cent of grades submitted by teachers were lowered by one full grade and 3 per cent were lowered by two grades following the moderation process.[88] Four days later in what was described as "a spectacular U-turn",[89] the Government scrapped the standardisation model – described by the then Prime Minister as a "mutant algorithm"[90] – for both A-level and GCSE results. This meant that the vast majority of students ended up receiving the grades their teachers had originally assigned to them (unless the algorithm had generated a higher grade for students, in which case they received whichever grade was highest).

As a result of the moderation process being abandoned and teachers' original grades standing in many cases, a dramatic increase in the top A level grades was clearly evident. The proportion of A* grades nearly doubled from 7.7 per cent to 14.3 per cent, and the proportion of A grades rose from 25.2 per cent to 38.1 per cent. In addition, C grades increased from 75.5 per cent to 87.5 per cent.[91] When GCSEs results were released two days later, a similar pattern

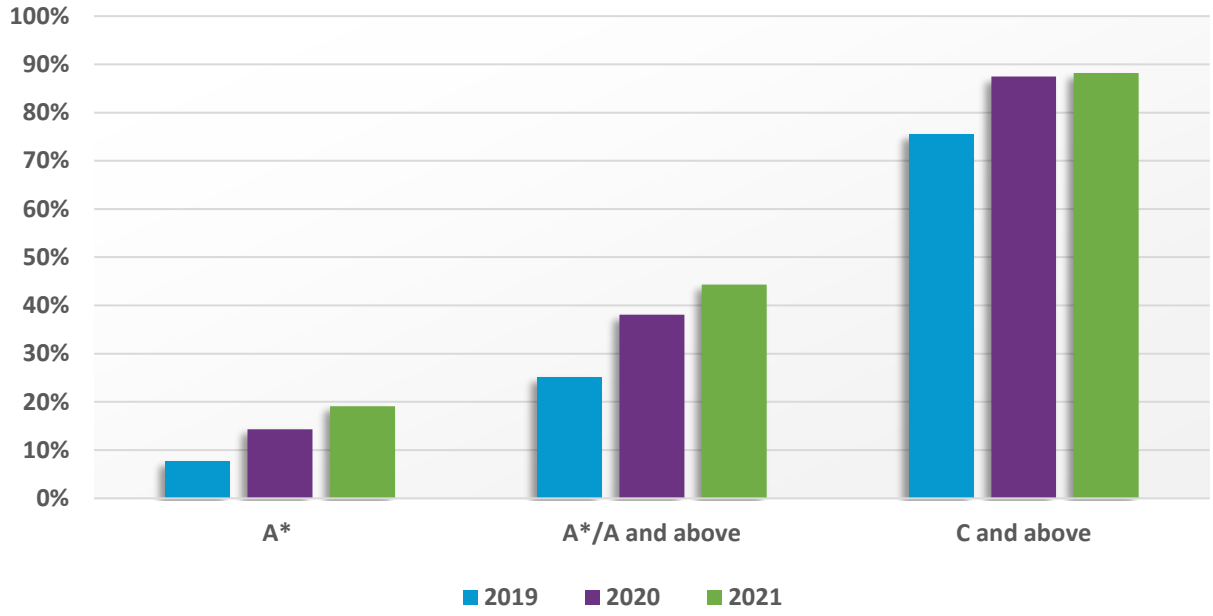
emerged. In 2019, 69.9 per cent of papers were rated grade 4 (previously grade C) or above but this rose to 78.8 per cent in 2020. The proportion of top grade 9s increased from 4.5 per cent to 6.3 per cent, while grade 7s (equivalent to an old A grade) or above rose from 20.6 to 25.9 per cent.[92] Unsurprisingly, the Government was keen to avoid another policy u-turn the following summer, particularly with the ongoing uncertainty about the COVID pandemic. As a result, they announced in January 2021 that summer exams would be scrapped for a second year running. The Government elaborated on their plans in February that year:

> *"For GCSEs, AS and A levels, teachers will assess the standard at which you are performing based only on what you have been taught so that your school or college can determine your grade. Teachers' judgements should be based on a range of evidence relating to the subject content that your teachers have delivered, either in the classroom or via remote learning. Teachers will be able to use evidence about your performance gathered throughout your course to inform their judgement. This might include work that you have already completed, mock exam results, homework or in-class tests. Your teachers may also use questions from exam boards, largely based on past papers, to help assess you, but this won't be compulsory."*[93]

Teachers were essentially asked to assess the standard pupils were working at, rather than the previous year's approach where they had to predict how well pupils might have done if they had taken the exam. There was to be no algorithm in 2021, although the Government announced that "exam boards will put in place quality assurance arrangements to make sure consistent judgements are being made",[94] with headteachers or principals signing off all grades. As shown in Figure 1 (overleaf), when the 2021 A-level results day arrived it was clear that grades had once again risen from 2020 to 2021, with both 2020 and 2021 representing a dramatic increase from pre-pandemic levels. When GCSE results were released, a similar albeit less dramatic increase was evident. The proportion of grade 7s and above had risen from 27.5 per cent in 2020 to 30.0 per cent in 2021. The proportion of grade 4s and above also rose to 79.1 per cent from 78.8 per cent in the previous year.[95]

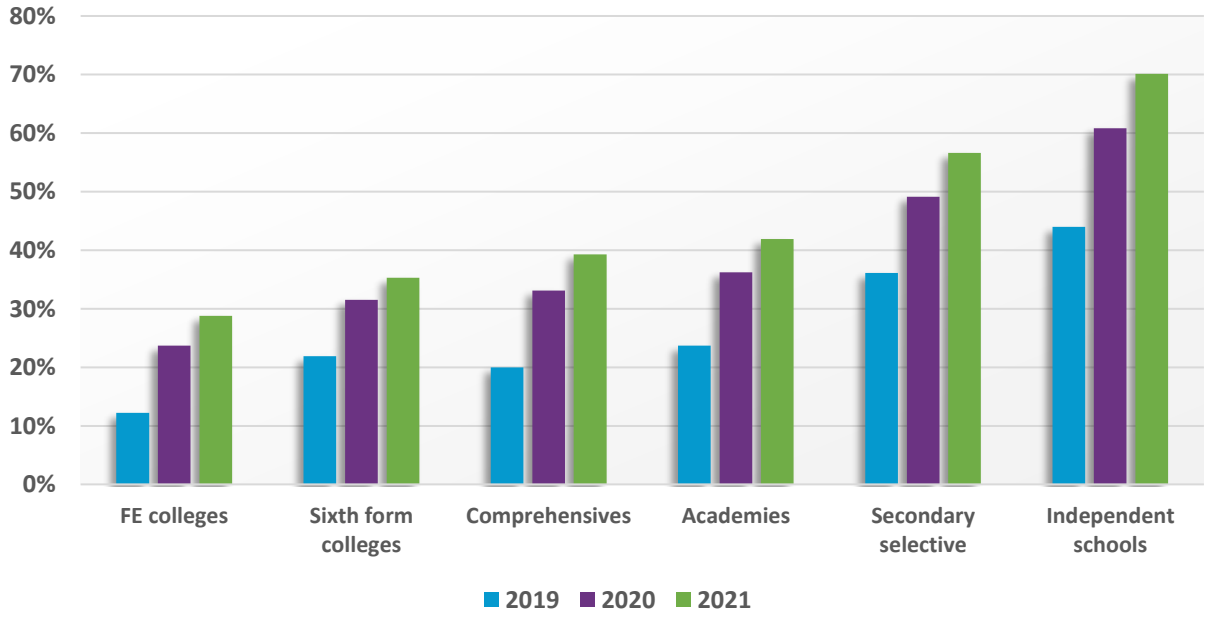
Ofqual's interim chief regulator Simon Lebus said he was "comfortable" with the 2021 results, with teachers having been in the "best position" to make judgements given the circumstances.[96] Even so, a subsequent survey by Ofqual of 3,000 stakeholders (including students, school and college leaders, teachers, parents and the general public) found that just 34 per cent of respondents had confidence in the reliability of A-level grades in 2021 compared to 72 per cent who reported having confidence in pre-pandemic grades. Meanwhile, confidence in the reliability of GCSE grades was 39 per cent in 2021 compared to 69 per cent before the pandemic.[97]

Figure 1: Percentage of students awarded different A-level grades [98]



In addition, concerns emerged that some students had benefitted more from teacher-assessed grades than others, with over 70 per cent of all A-level grades from private schools in England being awarded an A grade or higher compared to 60.8 per cent in 2020 and 44 per cent in 2019 when exams had last taken place. Figure 2 shows that although all school types saw an increase in the cumulative percentage of A grades awarded, independent schools saw the highest increase.

Figure 2: Cumulative percentage of A grades or above awarded by different types of providers [99]



GCSE grades followed a broadly equivalent pattern to A-levels. Overall, the cumulative percentage of grade 7s or above rose for comprehensive schools, academies, free schools and secondary selective schools by between 7.7 per cent and 10.5 per cent, but independent schools saw an increase of 14.6 per cent. Independent schools also saw a greater percentage point increase than secondary selective schools, despite secondary selective schools having a higher percentage of grade 7s or above in 2019 (57.9 per cent vs 46.6 per cent).[100] These large (and growing) performance gaps did not go unnoticed, with exam boards subsequently investigating allegations of malpractice in independent schools relating to teacher-assessed grades.[101]

Concerns over widening disparities between students

Not only were there gaps in attainment between types of schools during the pandemic, but there was also a widening of existing gaps based on demographic characteristics. Black pupils, pupils receiving free school meals (FSM) and pupils with a very high level of deprivation saw their performance relative to their reference groups (white pupils; non-FSM pupils; pupils with medium levels of deprivation) widen by 1.43, 1.42 and 1.39 percentage points respectively.[102] Similarly, the longstanding GCSE attainment gaps between FSM pupils and pupils with special educational needs and disabilities (SEND) relative to their prior-attainment-matched pupils widened by 2.27 and 2.00 respectively.[103] In response to these results, the Sutton Trust said that the COVID-19 crisis had "compounded existing inequalities"[104] and the Education Policy Institute said that the widening of the gaps for poor children and black children suggests that "this isn't the system that we would want in its entirety in the future."[105]

The COVID-19 pandemic was clearly a period of considerable disruption and teachers were put under significant pressure when asked to award grades to their students. That said, this was not the first time that disparities between students have arisen in the context of teacher-assessed grades. Indeed, there is strong evidence that teacher assessments are likely to be less valid and reliable than external tests due to the potential sources of 'bias':

- A 2016 meta-analysis on the existence of bias in grading students' work in Australia found that "statistically significant" rates of bias "can occur… when graders are aware of irrelevant information about the students'" including their racial/ethnic background, 'education-related deficiencies' and poor prior performance.[106]
- Burgess and Greaves (2009) found that teacher assessments "might be severely detrimental to the recorded achievements from poor families, and for children from some ethnic minorities", while "external testing in some way protects pupils from typically low-attaining groups from subconscious assumptions."[107]

- Using data for almost 5,000 pupils, research by Campbell (2015) "demonstrates biases in teachers' average ratings of sample pupils' reading and maths 'ability and attainment'" which corresponded to key demographic characteristics such as income, gender, special educational needs and ethnicity.[108]
- In a study of teacher bias in grading immigrants and native children in middle school, Alesina et al. (2018) found that teachers gave "lower grades to immigrant students compared to natives who have the same performance on standardized, blindly graded tests."[109]
- A study into teachers' attitudes towards students with high- and low-educated parents found that teachers showed "positive implicit attitudes towards students with highly educated parents", even though the teachers did not express different explicit (stated) beliefs regarding the learning and social behaviours of different students.[110]
- Doyle et al. (2022) found that "teachers judged students of lower [socioeconomic status] to be inferior to students of higher SES across a range of indicators" in relation to the quality of their work and their future potential.[111]

In light of concerns around the risk of bias in teacher assessments, a recent review by Ofqual found that the evidence on gender and ethnicity bias was mixed but the prospect of bias against pupils from disadvantaged backgrounds and pupils with SEND was much clearer.[112] These findings informed Ofqual's guidance to teachers during the pandemic that included practical suggestions for how to mitigate the risk of bias, such as reminding teachers to base each judgement purely upon evidence of how well a student has performed rather than other factors - for example, their attitude or behaviour.[113] It is unclear at this point how much of an effect such suggestions had in practice.

On the wider issue of the credibility of teacher assessment, a government-commissioned review in 2011 of teacher assessments in primary schools recognised that there were risks "that judgements will not be made consistently by teachers across the country".[114] The review concluded that "the evidence… does not suggest to us that moderation would address the considerable risks around reliability of moving to a system based entirely on teacher assessment."[115] Also in 2011, a summary of research in the USA found that "teacher classroom summative assessment, that is, teacher grading practices, have historically and currently emphasised the lack of validity and reliability of these judgements."[116] More recently, the 2017 Education Select Committee into primary school assessment in England "heard a wealth of evidence of the disadvantages of using teacher assessment within a high-stakes accountability system."[117] Tim Oates from Cambridge Assessment cautioned the Committee that "we have to be very realistic in terms of the level of dependability that we can yield from teacher assessment and whether it is always fair to expect teachers to assess with a level of consistency that we expect when we use the data for particular purposes."[118]

Leaving aside the issues of reliability and cost associated with teacher assessment, there is also a substantial workload burden attached to it - as exemplified by numerous media reports during the pandemic of the excessive workload generated by using teacher assessment in a high-stakes environment. During this period, it was common to hear teachers highlighting the number of 'late nights' required to get through the necessary marking as well as feeling their workload had 'gone through the roof' and led to sheer exhaustion.[119] What's more, if teacher assessment were to be used more widely, it would generate a number of practical issues such as the time and investment required to provide sufficient additional training for teachers and leaders in every school and college. In other words, there would be significant hurdles to overcome before teacher assessment could be feasibly implemented even if it was ever deemed appropriate within a high-stakes setting.

5. Oral assessments

An 'oral assessment' refers to "any assessment of student learning that is conducted, wholly or in part, by word of mouth."[120] There are many reasons for using oral assessments, including the opportunity to probe a student's knowledge, letting a candidate demonstrate their knowledge in a more practical form and helping to ensure academic integrity.[121] There are several options for how, why and when to use oral assessments:

1. **What is being assessed?** Oral assessments typically measure one or more of the following: knowledge and understanding; problem solving (particularly in novel situations); communication skills; and personal qualities such as confidence and professionalism.[122]
2. **How much interaction will there be?** Oral assessments can allow for interaction between the examiner/s and the student, and sometimes others (e.g. peers),[123] which can take different forms such as presentation or a dialogue.[124]
3. **How much does the assessment replicate 'real life'?** Oral assessments can seek to replicate a real life setting, such as OSCEs (Objective Structured Clinical Examinations) where nurses and other students of health professions are presented with 'patients' and discuss diagnoses and treatment plans with an examiner.[125] Alternatively, the assessment can be removed from such settings (e.g. a 'viva' for a doctoral thesis).[126]
4. **How structured will the assessment be?** Some oral assessments follow a closed structure (the examiner presents a pre-determined set of questions or events in a given order), whereas others follow an open structure with a looser agenda.[127]
5. **Who assesses it?** There are many possible assessors including authority-based assessment (e.g. an architect being a member of 'design jury'), peer assessment of a presentation or performance, and self-assessment (e.g. where students critically reflect on their own work and identify specific strengths and areas for improvement).[128]

Although oral assessments are considered standard practice in many professional and Higher Education contexts, they do not feature heavily in school or college assessments apart from language studies (see next section). This contrasts with countries such as Germany, where the *Abitur* (the qualification granted at the end of secondary education) incorporates oral exams lasting 20 minutes to assess at least one 'basic course' (as opposed to their main three 'intensive courses'). These oral exams consist of two components: a presentation (responding to a topic that they are given 30 minutes in advance of the exam) followed by a wider discussion.[129] In addition, an oral assessment can be used instead of a written exam to pass the *Abitur* if a student's written exam performance was insufficient in one of their main subjects.[130]

Oral assessment in schools

In England, assessments of oral skill are a central feature of modern foreign language qualifications. For example, GCSE French features an oral component worth 25 per cent of the qualification that assesses "communicating and interacting effectively in speech for a variety of purposes".[131] It consists of three sections: a role play (where students answer questions and convey information, as well as asking a question), a photo card (the teacher asks five prescribed questions, three of which will be printed on the student's card) and general conversation that involves asking and answering questions and exchanging opinions. The oral assessment lasts between 7-9 minutes for foundation students and 10-12 minutes for higher tier students. It is conducted and audio-recorded by the student's teacher but marked by an external examiner based on factors such as range and accuracy of language, pronunciation and intonation, spontaneity and fluency.

Two frequently cited studies explored the differences in marks given by teachers and moderators in GCSE French oral exams. The results showed that teachers were typically more generous than external moderators by about 3 to 5 marks depending on the difficulty of the oral assessment (equivalent to around 0.5-1 grade higher).[132] However, when teachers were asked to place candidates in rank order using the same marking criteria, they were nearly as effective as assistant examiners who mark conventional exam papers.[133] A study of the oral proficiency of US high school students in French and Spanish also found a significant difference between the teachers' chosen score for their students (mean 4.9) and the students' score from independent testers (mean 3.4).[134] Students with higher previous attainment were also overestimated by a greater amount, suggesting that teachers' judgements can be biased by knowledge of students' prior grades.[135]

Another tranche of research explored the reliability and validity of oral assessments. One study asked 24 teachers to mark recorded conversations of 30 higher-tier GCSE French candidates on two occasions one month apart. On each occasion, teachers had to make three different judgements on students:

A. A holistic judgement by allocating the student to one of four 'bands' using broad descriptors, and then selecting a high or low mark within the chosen band (creating a scoring range of 0 to 8 across the four bands);

B. Give individual marks for three categories - content, accuracy and pronunciation **-** on a scale of 0 to 3 based on detailed marking criteria for each category;

C. Give individual marks for three categories - fluency, range of vocabulary and complexity of structures - on a scale of 0 to 7 with no criteria or guidance.

Perhaps surprisingly, the approach that produced the poorest reliability between teachers was 'B' (marking against detailed criteria), with correlations (on a scale of 0 to 1) between the two marking exercises of .68 for content, .49 for accuracy, .46 for pronunciation. Even when the three categories were aggregated, the correlation between teachers was just .64.[136] In contrast, approach 'C' (marking on a scale with no criteria or guidance) produced correlations of .69 for fluency, .70 for range of vocabulary and .64 for complexity of structures, with a higher overall correlation of .73.[137] This increase in reliability in the absence of marking criteria occurred despite teachers who were interviewed after the experiment expressing serious reservations about 'C' precisely because of the lack of criteria. Another key finding was that approach 'A' (a holistic judgement of performance) produced a correlation of .69 between teachers, even though they only gave a single overall judgement about pupils' performance during the assessment without any marking guidance or criteria.[138]

The same study also compared the consistency of teachers in terms of their ranking of the students. After marking the same recordings on the two separate occasions, the consistency of the rankings by individual teachers ranged significantly from .93 to as little as .20.[139] Despite the inconsistency between individual markers, there were still high levels of consistency in the rankings in some areas e.g. the ranking of the holistic judgements in 'A' had a correlation of .82 between the two occasions[140] - providing another indication of teachers' expertise in ranking rather than grading students.

Oral assessments have previously been used more widely than in just language qualifications. For example, the Certificate of Achievement (CoA) was aimed at students aged 16 who were unlikely to achieve a grade in their GCSE exam (e.g. students with special educational needs). The CoA for Geography contained an oral assessment component alongside a written exam and coursework, which was described as "the first re-introduction of oral exams in schools (other than language exams) in Britain for 50 years."[141] The oral assessment was an audio-recorded interview lasting approximately 10 minutes, during which a student's teacher would ask set questions about resources from a booklet provided to the student in advance. A study in 1999 found that this oral exam had advantages over traditional written assessments, with teachers able to interact with students and prompt them to answer to the best of their ability.[142] Even so, the opportunity to prompt students decreased the reliability of the assessment as teachers were unable to be consistent with their prompts to different students.[143] In addition, teachers often had a 'right' answer in mind that students struggled to reach, which had the potential to cause students to become disaffected.[144]

Regardless, the researchers claimed that for students taking the CoA "the oral exam provides an opportunity for positive achievement which many of them may not find in traditional written exams" and "also allows us to gain a clearer idea of students' understanding of a subject".[145] Similarly, Schoultz et al. (2001) found that students scored better in oral physics

and chemistry tests than written ones, with qualitative analysis showing that students often failed to interpret the written questions without guidance. The authors concluded that the ability to re-phrase the question in an oral setting allowed a genuine test of students' conceptual understanding, thereby improving the validity of the test.[146]

Oral assessment in higher education ('viva voce examination')

A 'viva' is a compulsory element of being awarded a PhD in the United Kingdom, during which doctoral students are required to verbally defend their thesis in front of several internal and external examiners. The viva typically lasts for one to two hours, followed by a feedback session in which the examiners share the recommendation they have decided upon with the student (and their supervisor).[147]

Despite their widespread use, the variation in practice between institutions is widely acknowledged and can be unnerving for examiners, particularly those with limited experience. Some institutions have tried to increase the transparency and fairness of vivas by, for example, requiring an 'independent chair' whose role is to ensure 'fair play', requiring the viva to be digitally recorded to provide a clear record of events, and regulating the degree of existing relationships permissible between the proposed examiner and students.[148] Notwithstanding these attempts to improve transparency and fairness, the absence of a standardised approach across institutions makes some variation almost inevitable. In addition, there is variation in the perceived purpose of the viva (e.g. checking understanding to guard against malpractice in the thesis; clarifying problems or weaknesses in the thesis; deciding between borderline cases; testing a candidate's presentation skills).[149] Needless to say, examiners bringing their own interpretations of the purpose of the viva is likely to hinder its validity and potentially its reliability.

Despite their significance within the education sector, vivas are largely under-researched as an assessment tool.[150] Within the limited research that does exist, there are contrasting views about their value and reliability.[151] A recent study of 27 academics at 16 UK universities highlighted numerous issues through questionnaires and semi-structured interviews:

- Concerns about examiners' behaviour or attitudes, such as viewing the viva as a 'rite of passage' potentially being more likely to approach it with a confrontational spirit;
- Questions over whether the reliability of judgements, interpretation and subjectivity between examiners is due to "inconsistencies of approach";
- Doubts about whether cultural or emotional factors are taken into account by examiners e.g. a disinclination among some candidates to challenge authority figures;

- Uncertainty about the impact of 'cognitive style' on how a candidate is perceived by examiners e.g. a 'slow thinker' may not perform particularly well during the viva but may offer very perceptive observations if given sufficient time[152]

Other studies have identified further aspects of vivas that can threaten their reliability and validity. Torke et al. (2010) stated that vivas are appealing due to their high 'face validity' (appearing to be an effective form of assessment), flexibility and potential to measure aspects of competence that are not tapped by written exams. Nevertheless, they are prone to errors such as 'errors of contrast' (judgements of a candidate being influenced by impressions of preceding candidates) and 'halo effects' (a judgement of one attribute influencing judgements of others).[153] An example of a halo effect is that a relaxed, fairly eloquent but weak student may receive a better rating than their performance warrants compared to a knowledgeable student who has difficulty in expressing themselves.[154] It has also been suggested that vivas create an intimidating atmosphere that may worsen any anxiety and nervousness that a student is experiencing.[155] Knight et al. (2013) found that there can be high levels of anxiety but quoted other research that found no evidence that a viva was more stressful than other assessments.[156] In any case, Mellanby et al. (2011) found that students with higher anxiety scores on a self-reported questionnaire obtained the best (first class) degrees.[157] Overall, the evidence on viva-induced anxiety and its impact on performance is unclear.[158]

Oral assessment in professional certification

Oral assessment is frequently used for professional certification, and numerous studies have been conducted into how effective oral assessment is in a professional context. One common research topic is whether oral assessment alone is an effective indicator of attainment or whether it needs to be combined with other assessment methods ('multi-modal assessment'). Nunnick et al. (2010) found a moderate correlation of .58 between the written exam results of 45 senior trainees in intensive care medicine and their results from one of two 'live' assessment formats (either a simulation format or oral viva).[159] They concluded that this supported the use of multimodal assessment,[160] as it was unclear whether one mode is better at determining actual procedural competence or whether different modes of assessment assess different domains of performance.[161] When Torke et al. (2010) assessed the performance of medical students in their physiology theory written exams and vivas, they too found no consistent relationship between performance across the two assessment methods.[162]

Another area of research is the impact of a structured versus unstructured oral assessment. Mallick and Patel (2020) looked at medical students' and examiners' perceptions of structured versus unstructured vivas in biochemistry. The study found that many students felt a structured approach meant there would be less bias and it would induce less anxiety,[163] with

85 per cent preferring this approach.[164] Examiners also thought that a structured approach was fairer overall. Another study showed that by using a structured evaluation form for obstetrics/gynaecology students' oral assessments during clinical training, the reliability of examiners' marks increased from .48 to .67.[165] This indicates that by adding structure to the marking process, the reliability of scoring can be improved.

Similarly, Wass et al. (2003) found that the reliability of pass/fail decisions made by examiners for the oral component of the Royal College of General Practitioners' (MRCGP) membership could be improved if the oral exams were extended by 20 minutes (covering 15 topics instead of 10) with two further examiners (six versus four).[166] They concluded that the length of the oral exam and hence the breadth of topics covered had the greatest impact on its reliability. Wakeford et al. (1995) explored a range of errors that can occur in the MRCGP oral exam, including disagreements among examiners about grades, allowing first impressions to be overly influential and being influenced by a candidate's appearance. They suggested that extensive training is required to avoid these problems, as well as carefully selecting and monitoring examiners, planning each oral exam as a whole and having contingency plans for challenging candidates e.g. those who are less talkative and require further prompting.[167]

As noted earlier when discussing oral exams in schools, any use of prompts could undermine the reliability of the assessment given that examiners may use them in different ways for different students. This emphasises how, even in professional practice, these assessments still involve some trade-offs in terms of how they are designed and delivered (e.g. structured versus unstructured conversations; whether an examiner knows the candidate). Nevertheless, oral assessments have many commendable features – particularly in terms of building skills that written exams ignore (and are often prized by employers) such as verbal communication – while also testing some elements of knowledge and understanding. Well-designed oral assessments are therefore a strong contender for being able to make a greater contribution within a future high-stakes assessment system.

6. Portfolios

A 'portfolio' refers to "a collection of various forms of evidence of achievement of learning outcomes".[168] In practical terms, a portfolio for assessment purposes may include the following elements:

- **Evidence**: the type of evidence depends on the discipline or profession it is being produced for, but some examples include reports, papers and photographs.
- **Labelling of evidence**: evidence may be labelled to provide information on who contributed towards a specific piece, when it was produced and what it represents.
- **Structuring and signposting the portfolio:** portfolios can be sizeable, meaning a clear and explicit structure is vital for both the creator and the assessor. Portfolios can be structured in many ways (e.g. by natural topic headings, or chronologically).
- **Critical reflection or commentary**: students can contextualise the evidence in their portfolio by, for example, explaining what it shows about what they have learned and their current capability and understanding of a subject.[169]

Portfolio assessment in schools

Portfolio assessments are currently used in some GCSEs and A-levels. For example, they are a significant feature of the Design and Technology GCSE specification, which contains a portfolio 'non-exam assessment' component worth 50 per cent of the final grade. Students are expected to "generate design ideas…and develop these to create a final design solution (including modelling)"[170] as well as creating a final prototype. Students must evidence, investigate, analyse and evaluate their work throughout a portfolio of approximately 20 pages of A3 paper (or the A4 / digital equivalent).[171] Portfolios are completed under direct supervision and are marked at the end of the course by the students' teacher using guidance provided by the exam board. A sample of portfolios are then sent to the exam board for moderation, which involves a moderator re-marking a sample of the evidence and checking that the marks given by the teacher are in line with the agreed standards.[172]

One of the most important advantages of portfolios is that they are more closely associated with real-world situations than traditional paper-and-pencil tests. Portfolios also allow students to demonstrate "evidence of growth over time" through gathering extensive information about how they think and reason, how they apply data to solve problems and how responsive their work is to feedback.[173] What's more, portfolios encourage students to be responsible for their learning, to critique their own work, to make connections and extend

their learning – thus allowing them to demonstrate "skills more complex than those required to recite facts".[174] Other potential benefits include improvements in students' knowledge and understanding as well as improved student-tutor relationships.[175]

Some of the most widely cited studies of portfolios were conducted by Professor Daniel Koretz and colleagues, who analysed their usage in elementary and high schools in the US over several years. In Vermont, state-wide portfolio assessments were introduced for mathematics and writing in the fourth and eighth grades.[176] Students were required to maintain year-long collections of a range of their work in both subjects, which were scored on multiple dimensions.[177] Scoring for state-wide reporting was normally carried out by a teacher other than the students' own.[178] It was found that the consistency between assessors ('raters') was typically lower in writing than in mathematics.[179] In the first year of the programme, correlations between the scores awarded by the raters ranged from just 0.38 to 0.49 for both subjects.[180] Over the years, raters became more consistent due to refinements to the scoring rubrics and further training. As a result, correlations between raters in mathematics eventually reached 0.8-0.9 but the correlations in writing remained lower at 0.64-0.66.[181]

Portfolio assessments for writing in the fourth, eighth and twelfth grades in Kentucky achieved similar levels of consistency (reliability). Raters were only required to provide a single score for a student's entire portfolio,[182] which produced an inter-rater reliability score of around 0.7.[183] A separate trial of writing portfolio assessment by LeMahieu et al. (1995) in Pittsburgh had raters score the portfolios holistically on three dimensions, which produced inter-rater reliability of 0.76-0.87.[184] Another trial in 1990 by the National Assessment of Educational Progress (NAEP) also invited a sample of fourth and eighth graders to produce a writing portfolio. The results were promising, with correlations between raters of 0.79-0.89,[185] although an almost identical trial two years later saw the correlation drop to 0.59-0.68.[186]

Research studies in the UK have faced similar struggles to achieve high reliability for portfolios. Wolfe (1996) reported on a nationwide portfolio pilot in which over 2,000 secondary students submitted portfolios from language arts, mathematics and science classes. Inter-rater reliability for mathematics and language arts was found to be 'satisfactory' at 0.66 and 0.62 respectively,[187] while the reliability for science was 'lower than would be desirable' at 0.44.[188] However, it was suggested that adequate reliability could be achieved by two raters scoring multiple portfolio entries.[189] Such are the difficulties inherent in assessing portfolios, a 1998 review of National Vocational Qualifications even concluded that it is impossible to develop written descriptions that are so tight they can be applied reliably to portfolios by multiple assessors in multiple assessment situations.[190] The review suggested that there may be an 'optimal degree of precision' for descriptions, as too much detail makes the portfolio unworkable in practice while too little causes the whole process to lack focus, yet this optimum is difficult to identify.[191]

Artistic portfolios may present even greater barriers to achieving consistency between scorers. Myford and Mislevy (1995) analysed data from art portfolios submitted to support US college entrance, in which each portfolio received 13 ratings (e.g. colour, design) as well as being summarised into an overall score. The individual ratings produced reliability scores of around 0.65-0.7, with the overall score producing a reliability of 0.78.[192] When questioned afterwards, raters offered a variety of reasons for why judging the portfolios was challenging, including:

- The 'bounce effect' (after rating one portfolio highly, the next portfolio may receive a lower score);
- Resisting operating in 'empathy mode' (where the rater judges a student's potential rather than the portfolio itself);
- Lacking background and experience in the medium or style in which the student is working; or, conversely, having a considerable background and experience in the relevant medium or style;
- Portfolios where the student has good ideas but does not have the technical capabilities to see the ideas through;
- Portfolios that depart from 'the norm' (i.e. are eccentric or daring), employ unusual media or are based on non-Western traditions;
- Uneven quality within the portfolio, with some stronger and weaker material.[193]

These challenges partly explain why it took six days to complete the judging process for several thousand portfolios, with the authors recognising that "performance assessments demand more resources than multiple-choice assessments".[194]

Alongside the reliability of portfolios, some research has explored its validity as an assessment tool. Leon and Elias (1998) compared scores in portfolios, 'performances' (e.g. observations, experiments) and traditional assessments in middle school and found that students' marks differed depending on the assessment method used. Some students did better in the portfolio assessment – particularly those who were generally lower achievers – indicating that this method could be capturing different elements of attainment than traditional assessments. As a result, portfolios could be considered a valid form of assessment depending on what it is intending to measure. Koretz (1998) found that the correlations between pupils' portfolio scores in mathematics and writing and the same pupils' scores on standardised tests were around 0.35,[195] leading him to suggest that 'moderately high' correlations between two measures of the same subject may "provide the best evidence of validity"[196] because the measures were capturing different attributes. Even so, he noted that "one would expect that scores on the mathematics portfolios would correlate more strongly with the mathematics uniform test than with the writing uniform test [but] this was not the case."[197] In the end, Koretz concluded that "the optimal level of association among measures …remains arguable" and "in general, the evidence pertaining to validity [for portfolios] was unpersuasive."[198]

Other concerns with the validity of portfolios include the question of who assesses them. In the Kentucky portfolio study, Koretz observed "a sizeable difference in the standards applied by students' own teachers" with mean scores dropping 'substantially' when the portfolios were rescored.[199] This problem of "overly lenient scoring by classroom teachers" appeared to reduce over several years of the portfolio programme but evidence on the extent of the improvement was "not entirely clear".[200] Another validity issue identified in Kentucky was the variation in the amount of support provided to the student by the teacher. A large majority of teachers reported varying the amount of assistance they provided "in response to the needs of the students", which Koretz concluded "calls into question the comparability of the resulting products and scores".[201]

Moreover, the time-consuming task of assessing portfolios could pose a threat to their validity, reliability and practicality. Supovitz et al. (1997) looked at portfolio assessment in New York primary schools and, similar to previous studies, the reliability scores recorded for the external portfolio raters was no higher than 0.68-0.73.[202] This mediocre reliability was associated with a lack of material within the portfolios, yet the raters explained that it would have been too time consuming for the portfolios to cover all the work required in sufficient detail for them to be able to properly rate it.[203] Koretz had also commented that even in programmes where portfolio assessment has been successful, it must be weighed against its limitations "such as the considerable costs in time, money and stress that this type of assessment entails."[204]

Aside from the workload implications, fundamental concerns over the suitability of portfolio assessment in high stakes environments are prevalent in the research literature. In line with the findings cited earlier from the large Kentucky study, Herman et al. (1993) cautioned that within a high-stakes setting "the pressures to increase instructional support in portfolio support may be substantial."[205] Meanwhile, the research by Supovitz et al. cited above argued that the reliability of portfolio assessment is too low for high stakes accountability purposes. Koretz recognised that portfolios are widely used for internal assessment "and for that purpose they might be more successful", but the evidence from large scale portfolio assessments was "not encouraging" as it often failed to overcome "one of the most basic and essential procedural hurdles"[206] of obtaining consistent scoring of student work:

> *"Portfolio assessment has attributes that make it particularly appealing to those who wish to use assessment to encourage richer instruction—for example, the 'authentic' nature of some tasks, the reliance on large tasks, the lack of standardization, and the close integration of assessment with instruction. But some of these attributes may undermine the ability of the assessments to provide performance data of comparable meaning across large numbers of schools."*[207]

Portfolio assessment in professional practice

Some professions make extensive use of portfolio assessments, particularly for training or revalidation purposes. As in school settings, the relevance of portfolios to real-world and professional environments is often praised. Elton and Johnson (2002) suggest that portfolios can offer an 'authentic' assessment that is "likely to provide predictive information" about how a candidate will perform in future in a professional context.[208] Portfolios have also been described as having 'high face validity', in that they appear effective in measuring the abilities that they seek to capture[209] and can provide a "holistic picture of the candidate".[210] On a broader note, it is claimed that portfolios help students to cope with uncertain or emotionally demanding situations as well as preparing them for the professional setting.[211] Other purported benefits are similar to those identified in school contexts such as improvements in a student's knowledge and understanding (particularly the ability to integrate theory and practice), greater self-awareness and reflection and the ability to learn independently.[212]

Baume and Yorke (2000) looked at the use of portfolios in the accreditation of higher education (HE) teachers in the UK at the Open University. The portfolios contained evidence (e.g. lesson plans, graded student work) to underpin the assessment of whether the course participant had met the desired outcomes.[213] The portfolios were doubled-marked across seven outcomes,[214] with a third assessor brought in where there was disagreement on scores.[215] Inter-rater correlations ranged from -0.10 to 0.67, with a median of just 0.24.[216] Agreement between assessors was especially problematic across some outcomes, with assessors often having quite different notions of what constituted an acceptable level of performance.[217] Similar findings were recorded by Centra (1994) for portfolios in American HE when evaluating faculty members in relation to contract renewal / promotion. Greater variation was observed in raters' judgements on specific elements of the portfolios such as the assessment of 'teaching' and 'service' compared to their judgements on more factual elements such as 'credentials' and 'participation in professional associations'.[218]

Another profession where portfolios are often used for accreditation is healthcare, yet problems with inter-rater reliability are evident here as well. A small-scale study by Pitts et al. (2002), which looked at the reliability of scores awarded to portfolios produced by a cohort of prospective GP trainers, achieved 'poor to moderate' inter-rater reliability of 0.1-0.41, which only increased to 0.5 with marking criteria discussions among the raters.[219] On a slightly more positive note, a systematic review in 2007 of portfolios in medical education found an 'average' reliability of 0.63[220] while a follow-up study in 2009 also obtained reliability of 0.87 for medical internship portfolios that had been double-marked.[221] That multiple assessors are often required to produce a reliable mark for a portfolio was echoed in other healthcare research (e.g. Jasper and Fulton, 2005[222]; Melville et al., 2004[223]), yet the practical implications of such an approach are obvious enough.

The struggle to achieve sufficiently high reliability for portfolios is compounded by potential problems with their validity in professional settings. Smith and Tillema (2001) looked at portfolios among different professionals including senior nurses and nursing staff and felt that evidence in portfolios had 'questionable validity' especially when used for assessment purposes.[224] In addition, Carraccio and Englander (2004) conducted a literature review on portfolio assessment in medicine and reported difficulties in striking a balance between the creative, reflective aspects of the portfolio (which is focused on learners) and a structure that is reliable and valid.[225] Another issue is ensuring that assessors do not base their marks on components that the portfolio is not intended to measure – for example, a perfectly competent teacher may struggle to render their practice into words and be penalised for this. Meanwhile, a 'bad' teacher may be able to produce a 'good' portfolio.[226]

Even the real-world relevance of portfolios – often seen as one of their greatest assets in a professional context – is sometimes called into doubt. Maidment et al. (2006) criticised the use of portfolios to meet dental professional body requirements for revalidation on the basis that "[a portfolio] doesn't prove you are a good or safe dentist, it proves you can fill a book".[227] Stocks et al. (2009) also argued that it is often challenging to determine "whether the portfolio presents an authentic experience, or is simply an effort to play the assessment 'game'".[228] Citing the work of Buckridge (2008), they noted that portfolios use for summative assessment are "likely to be preoccupied with demonstrating competence, with a focus on success", which limits the developmental potential of portfolios because "the point of the text is to persuade" and therefore "the mechanism loses authenticity".[229]

Given these challenges, portfolios may be more useful when utilised alongside other methods of assessment. For example, while Maidment et al. (2006) expressed concern about portfolios being used for revalidating dentists' fitness to practice, they accepted that portfolios could be beneficial if used as a basis of an appraisal interview.[230] The above research by Melville et al. (2004) also accepted that portfolio assessment for paediatric Specialist Registrars "had a place" as part of a triangulation process with other methods of assessment.[231] Jarvis et al. (2004) looked at portfolios by psychiatry residents in the USA and found that portfolios were unable to represent all six of the general competencies required by the Accreditation Council for Graduate Medical Education, and that it was therefore "reasonable and realistic" to use more than one form of evaluation method to examine performance.[232] Similarly, Van der Vleuten (2012) argued that when considering the clinical competence of medical students "there is no single bullet that can do it all in one go"[233] and that high-stakes decisions should be based on data from multiple methods of assessment.[234]

As noted earlier in this chapter in the context of school assessment, it is possible that portfolios could measure different elements of ability to written exams. Research by Davis (2001) investigated the level of agreement between ratings given to portfolios and other methods of

assessment for medical students. The correlation between the portfolio scores and an extended multiple-choice exam was 0.42, while the correlation between portfolios and a student's OSCE was 0.47.[235] It was concluded that rather than portfolio assessment being inherently unreliable, it was instead measuring both common and different abilities from those being tested in other forms of assessment[236] – as suggested by Leon and Elias for school-based portfolios. That said, another similarity between portfolios in schools and professional settings is the associated workload. Research by Hrisos et al. (2008) found that two thirds of foundation doctor trainees viewed a portfolio as a "burden", as the collection of the required paperwork was difficult to manage in busy hospital wards.[237]

In the end, as was found in school settings, the poor reliability and dubious practicality of portfolios means that their credibility as a large-scale assessment method within a high stakes system is weakened. Roberts et al. (2002) observed that while concerns around high reliability and validity may not be considered critical for portfolios used for formative purposes, these features become essential as soon as portfolios are used in high stakes decisions.[238] Davis et al. (2001) found that some final year medical students who completed a portfolio rather than traditional final exams even felt they had missed out on a 'rite of passage' and that they were less prepared for their junior doctor training as a result.[239] As discussed earlier, some researchers have proposed different ways of improving the suitability of portfolios for high-stakes judgements such as the use of triangulation between assessment methods. Nevertheless, as Koretz and others observed in relation to portfolios in schools, Roberts et al. (2002) concluded that "there is little evidence at present to support the widespread introduction of portfolios for high stakes summative assessment" in medical settings.[240]

7. Extended essays and projects

An alternative method of assessment that is sometimes found operating alongside written exams is the notion of an 'extended project'. Generally, such projects involve a student conducting independent research into an area of personal interest and then producing a final product (e.g. essay, physical artefact, live performance). This chapter will explore two widely recognised versions of 'extended projects' as a form of assessment.

The Extended Project Qualification (EPQ)

The 2005 White Paper by the then Department for Education and Skills aimed to "strengthen" A-levels by encouraging "greater stretch and challenge", including the development of an 'extended project'. This project would be "a single piece of work, requiring a high degree of planning, preparation, research and autonomous working" and, although the projects were expected to differ by subject, "all will require persistence over time and research skills to explore a subject independently in real depth."[241]

Following a successful pilot of the 'extended project qualification' (EPQ) in 2006, it was officially introduced in 2007/2008 as a voluntary qualification available for students in Key Stage 5 (ages 16 to 19) that can be taken alongside other academic and vocational qualifications. The QCA stated that the EPQ was intended to:

- Add depth and breadth to the curriculum;
- Allow learners to draw connections between subjects;
- Support learner progression and create links with their future study or employment interests;
- Help learners to develop new and enhance existing skills;
- Increase student confidence, responsibility and motivation.[242]

The EPQ is now recognised by universities and employers, with the highest grade EPQs being worth 28 UCAS Tariff points (equivalent to half an A-level).[243] According to UCAS, "the skills that students develop as part of the EPQ are highly valued" by higher education (HE) providers.[244] What's more, many universities lower their offer to applicants who are completing the EPQ, including some of the Russell Group universities such as Queen Mary University of London and the University of Manchester.

To complete the EPQ, students are required to:

- Choose an area of interest
- Draft a title and aims of the project for formal approval
- Plan, research and carry out the project
- Submit their research – the final output may be in the form of a long written report (approx. 5,000 words) or as an artefact e.g. a short film or an art piece, which must be accompanied with a short report (approx. 1,500 words).
- Deliver a presentation about their research
- Provide evidence in a log of all stages of planning and progress of the research, including decision-making and reflections on the process.[245]

Each student is allocated a supervisor, who they meet with for regular reviews over the course of the EPQ. This same supervisor marks the final project using the exam board assessment objectives and mark scheme. Supervisors must also confirm a presentation took place, and endorse the student's production log by signing a declaration that the evidence submitted is the unaided work of the student.[246] Supervisors send over all their students' marks to the exam board as well as a sample of EPQ submitted by students. A moderator from the exam board will then re-mark the sample of projects, and compare this with the marks provided by the supervisor to see if any changes are required.[247]

The EPQ has grown in popularity from just under 2,000 in its first year[248] to almost 40,000 entries in recent years, albeit with a slight dip in 2020.[249] Most students take A-levels alongside their EPQ (95.4 per cent), while around 9 per cent also took a BTEC qualification.[250] Furthermore, EPQ students were found to have had very slightly higher prior attainment than A-level students as a whole.[251]

Some research has suggested that the EPQ is associated with students' overall performance at Key Stage 5. After accounting for prior attainment and other background characteristics, Gill (2016) found there was a small but statistically significant effect, with those taking EPQ achieving better results on average in their A-levels than non-EPQ students[252] – equivalent to an uplift of one grade in one qualification if taking four A-levels.[253] Meanwhile, Jones (2016) found that after controlling for other variables, taking the EPQ enhances the odds of achieving a higher grade A-level (A*-B) by 29 per cent.[254] Interestingly, this positive impact was not uniform across A-level subjects, with no effects being found in mathematics or languages.[255] Surridge et al. (2021) reported similar findings: students who completed the EPQ had on average a 22 per cent higher chance of achieving grade A*-B at A-level than a peer who did not.[256] However, Jones (2016) recognised that looking for impacts of the EPQ on academic performance should be done cautiously, as it is possible the EPQ may act as a proxy for a

confounding variable – for example, EPQ students potentially being more motivated and thus performing better in their other subjects.[257]

In addition to the potential impact on other subjects, some research suggests the EPQ is better able to prepare students for HE. Gill (2022) found that EPQ students were more likely to progress to HE (88.5 per cent) compared to non-EPQ students (66.8 per cent).[258] What's more, first year university students who took the EPQ were less likely to drop out during their first year (2.3 per cent) compared to those who did not (5 per cent).[259] The same study also found that EPQ students were more likely to achieve a good degree (31 per cent achieved a 'First' and 87.7 per cent achieved at least a 2:1) than non EPQ students (24.6 per cent and 79.6 per cent respectively).[260] That said, previous research by Gill and Roderio (2014) found that the effect of having done an EPQ on the probability of achieving a First was not significant for Russell Group students but was significant for non-Russell Group students.[261]

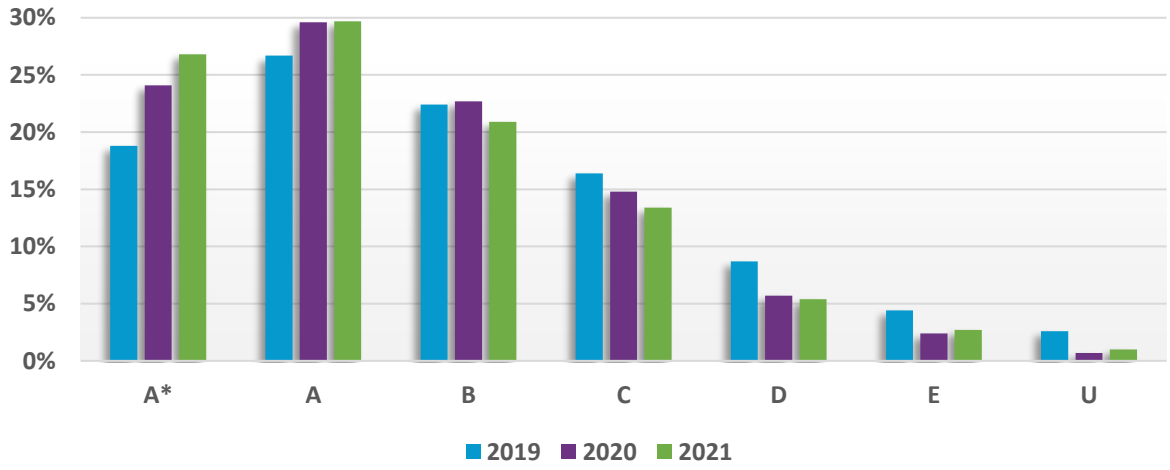
To explore why EPQ students may perform better in other A-level subjects and at degree level, Stephenson and Isaacs (2019) conducted focus groups with 26 teachers who had experience supervising or coordinating the EPQ and also interviewed 15 undergraduate students who had recently completed the EPQ.[262] The study reported that the EPQ influenced students' general academic performance by "enabling multiple opportunities for enhancing learning beyond the A-level curriculum by promoting self-regulation".[263] For example, both teachers and students perceived the EPQ as a way of empowering students to be 'agents in their own learning', and of improving their self-confidence and attitudes towards learning.[264] What's more, throughout the EPQ process learners were seen to make "discoveries" about themselves (e.g. their aptitude and their learning preferences), which was felt to enable students to 'optimise' their approach to learning.[265] The qualification was also perceived as a "catalyst" for engaging learners because the opportunity to pursue their own interests made the process enjoyable and improved their motivation.[266]

As described in earlier chapters, a significant proportion of GCSE and A-level coursework has been replaced by 'non-exam assessment' (NEA) due "largely to issues with authenticity and the reliability of marking" of coursework.[267] The use and weighting of NEA has intentionally been kept to a minimum by the exam regulator Ofqual, and where it is used, "significant new controls" are in place.[268] However, the EPQ is unusual as it is still assessed entirely through NEA by teachers. Ofqual has accepted that due to the nature of the EPQ and the wide range of potential projects that a student may produce, "the nature and level of regulatory controls that can be imposed are limited".[269]

In line with the grade inflation associated with teacher assessment and coursework discussed earlier in this report, Ofqual identified "modest" inflation in EPQ grades, with students from 2014 onwards generally having better EPQ outcomes than similar students from 2013 and

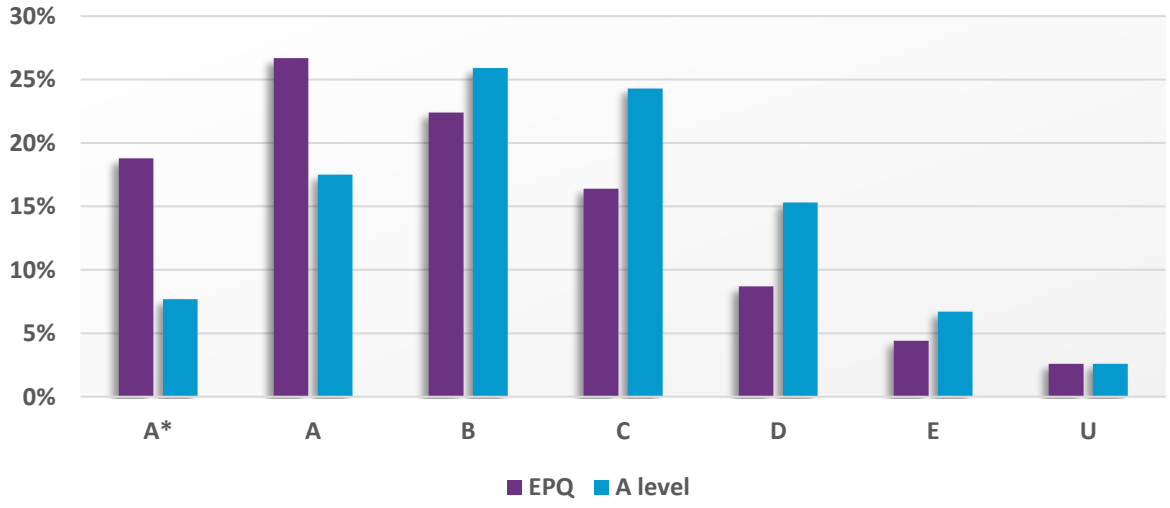
before. Analysis suggested that a grade C in 2016 was about half a grade easier to obtain than in 2010, and a grade A is a third of a grade easier.[270] As shown in Figure 3 below, there are further signs of grade inflation in recent years. In 2019, around 19 per cent of EPQ students received the highest grade (A*). However, during the pandemic this increased to around 27 per cent in 2021. What's more, the proportion of students awarded the lowest grade (U) reduced from 2.6 per cent in 2019 to under 1 per cent in 2020 and 2021. These pandemic-era figures highlight yet again the lack of controls on teacher-led assessment for the EPQ.

Figure 3: Percentage of different EPQ grades awarded [271]



Looking at pre-pandemic data from 2019, Figure 4 shows a stark difference when comparing the grades awarded to students in the EPQ and all A-level subjects, with far more top grades (grade A and above) being awarded for the EPQ. The percentage of students achieving an A or A* in the EPQ was 45.5 per cent in 2019 compared to just 25.2 per cent across all A-levels.

Figure 4: Percentage of grades awarded for EPQs and all A-levels in 2019 [272]



As discussed in the 'teacher assessment' chapter, there was a dramatic increase in top A-level grades during the pandemic, yet in 2020 and 2021 the proportion of top grades awarded for the EPQ remained 12-15 per percentage points higher than for all A-levels.[273] This suggests that the EPQ generates a similar 'bunching' of marks at the top end of the grade distribution – a persistent problem with coursework in previous years before Ofqual's new NEA rules were introduced. With such a high percentage of top grades, it becomes harder to view the EPQ grades as a credible assessment method within a high stakes assessment system.

The International Baccalaureate 'Extended Essay' (EE)

The International Baccalaureate (IB) Diploma programme is widely regarded as an academic alternative to A-levels in this country and abroad. UK colleges and universities accept the IB Diploma for entry onto most undergraduate courses. It is offered by around 100 schools in the UK, and in 2022, there were 5,250 IB diploma candidates across the UK.[274]

The IB has a much broader curriculum than A-levels in terms of subjects and tasks. The Diploma curriculum is made up of a 'core' plus six subject groups (e.g. Sciences; The Arts). The 'core' aims to broaden students' educational experience and challenge them to apply their knowledge and skills. It consists of three elements:

- **Theory of knowledge**, in which students reflect on the nature of knowledge and on how we know what we claim to know;
- **An extended essay,** an independent self-directed piece of research with a 4,000-word final paper;
- **Creativity, activity, service** - in which students are expected to take part in a range of experiences, and complete at least one project related to either 'creativity', 'activity' or 'service'. [275]

The Diploma is awarded to students who gain at least 24 points overall for the IB (with the highest possible score being 45 points), subject to minimum levels of performance including successful completion of the three elements of the 'core'. The 'theory of knowledge' and 'extended essay' (EE) components are awarded individual grades and are collectively worth a maximum of 3 IB points, depending on the grades obtained overall in each course.[276]

For the mandatory EE, students are expected to investigate a topic of personal interest to them, which relates to one of their six subject groups or takes an interdisciplinary approach. Students are "supported throughout" the research and writing of their extended essay, receiving "advice and guidance" from a supervisor who is usually a teacher at school.[277] Students are also required to have three 'reflection sessions' with their supervisor, with the

last of these acting as a concluding interview ('viva voce') in which students are asked to reflect on the strengths of their work and findings as well as any areas that caused problems and what can be learned from the report.[278] All essays are externally marked by examiners appointed by the IB.[279]

The IB states that by completing the EE, students develop skills in:[280]

- Formulating an appropriate research question
- Engaging in a personal exploration of the topic
- Communicating ideas
- Developing an argument
- Developing the capacity to analyse, synthesize and evaluate

The IB also claims that the skills provided by the EE provide "practical preparation for undergraduate research".[281]

Research has explored the perceptions of the EE among IB students in helping them to prepare for university. In a study conducted on behalf of the IB, Aulls et al. (2013) found that a large percentage of students who completed the EE reported "enhanced organization, reading and reasoning skills".[282] What's more, students "almost unanimously" mentioned that doing the EE "improved their confidence in their ability to accomplish the academic demands of undergraduate studies".[283] Similarly, Taylor and Porath (2006) found that graduates of the IB reported "positive experiences" and felt it "prepared them well for postsecondary studies",[284] while Inkelas et al. (2012) heard from former IB students that the EE helped with "the reduction of anxiety around college writing assignments".[285]

The study by Inkelas et al. also asked former IB students how well they felt their EE prepared them for college-level (or, in the UK, university-level) work. They were asked to rank 12 research skills (e.g. creating a reference list, analysing data) on a scale from 0-10, with 10 meaning that the EE gave them 'complete preparation'. As shown in Figure 5 (overleaf), most students selected 8 out of 10 or higher for 'preparation' in almost every skill. The study concluded that IB students felt the EE prepared them well for HE assignments and that it covered most of the skills needed for a future research project.[286] However, when this data on 'perceptions of preparedness' was correlated with students' current level of confidence with research skills, the correlations were between .24 and .47. This modest relationship suggests that students' confidence with research skills was only partially explained by their perceived level of preparedness due to their EE experience, and were instead explained "to a larger extent" by other factors in their backgrounds or college experiences.[287]

Figure 5: IB alumni's perceptions of how well the extended essay prepared them for college-level work for a variety of research skills [288]

Further research into the perceptions of the EE in preparing students for HE was conducted by Wray (2013). Semi-structured interviews were held with 24 former IB students as well as four focus group meetings with 18 students. Overall, students who had completed an EE were "very positive about their experience".[289] Many also noted that they had learnt a great deal from the experience, particularly study skills.[290] Wray found that some students enjoyed and benefitted from the independence required for the EE, but others struggled with the lack of support. One interviewee said that "I did a questionnaire but only found out they were bad questions at the end. I wish someone had told me."[291] What's more, most interviewees highlighted the difficulty in setting an appropriate research question and designing a project to answer this question sensibly.[292] This mirrors research by Hamer (2010), who found that students perceived the EE as "a cognitively complex task", and the ability to construct a question to guide their research was ranked the most problematic of the skills required.[293] Unlike the EPQ, the EE is compulsory for IB students so there is likely to be more variation in how well suited they are to tackling such complex tasks.

In addition to surveying the perceptions of students about the EE, some research has looked at whether there is a correlation between the EE and increased performance in HE courses. The research by Inkelas et al. (2012) found that students' EE scores were able to significantly predict their first semester college GPA, but the EE score only explained 1 per cent of the variance (rising to 4 per cent for their final semester GPA scores) after controlling for student

background characteristics.[294] The study concluded that the correlation between EE scores and cumulative college GPA was "modest at best"[295] – which echoes the research into EPQs and HE performance discussed earlier in this chapter.

Could extended projects prove useful in a high stakes system?

There are several notable similarities between the EPQ and the IB Extended Essay:

- Students are expected to conduct an independent research task from start to finish, which aims to develop their ability to analyse, synthesise and evaluate information;
- Students are able to choose a research topic that appeals to them personally, which encourages them to take responsibility for their own work;
- Students have regular check-ins with supervisors to discuss their progress and keep track of their development e.g. time management and organisational skills;
- Students must present their ideas at the end of the course and discuss the strengths of their research as well as the challenges they faced during the process.

These attributes and skills are quite different to those assessed by written exams, yet they are nevertheless thought to be important and valuable to instil in students. One of the clearest benefits of extended projects is that they develop skills that are required in real-world settings and explicitly set out to prepare students for life outside of school or college. Many of the skills required by extended projects are essential in both HE and the workplace, such as project management, communicating ideas and working independently. Another benefit of extended projects is that they are a credible way of students acquiring these skills through conducting the research themselves rather than being taught about it in a classroom. Moreover, these projects seem to be a valid form of assessment, with both the EPQ and the EE appearing to measure what they claim to – that is, a students' development of independent research skills.

Despite the numerous benefits of extended projects, their limitations are hard to ignore. Both the EPQ and the EE can be affected by a range of issues that correspond to the problems faced by several assessment methods discussed earlier in this report - in particular, coursework and other forms of teacher assessment:

- Problems with determining the authenticity of students' final submission and whether it has genuinely been completed independently;
- Teachers providing too much assistance to students during the project, thereby influencing their final grade and reducing the validity of the assessment;
- Inconsistencies across teachers and schools in the way that the required controls and guidance are implemented;

- The risk of bias in the assessment process, which could affect the reliability and validity of grades (especially the EPQ, which is marked by a student's supervisor)
- The pressure on teachers to award higher grades in a high-stakes system (as highlighted by the bunching of top marks in the EPQ).

Confirming that a student has completed their project independently has become increasingly problematic. With extended projects, students are trusted to act with honesty and integrity during the research process. Although there are some checks in place to encourage students to focus on their learning and development – such as the EPQ presentation and the EE viva – these are unlikely to catch or prevent malpractice on the part of the student. With tens of thousands of extended projects being submitted every year, it is highly unlikely that safeguards which are both practical and effective could be implemented.

These existing concerns are now being exacerbated by the development of ChatGPT and other chatbots. In February 2023 – just three months after the public release of ChatGPT – a cheating scandal erupted at a high school in Florida, USA, after several students admitted to using ChatGPT or other AI software to write their EE, while other students were suspected of having done the same but had not admitted as much.[296] In the same month, the IB announced that in future students will be allowed to use content generated by ChatGPT so long as they credit ChatGPT as the source and do not pass it off as their own – a situation made even more remarkable by the fact that only two months earlier in December 2022, an IB spokesman stated that "the use of ChatGPT or any other AI tool is against the IB's academic integrity policy", adding that "academic integrity is an essential aspect of teaching and learning in IB programmes where the action is based on inquiry and reflection".[297] The head of assessment principles and practice at the IB recently admitted that essay-writing is "being profoundly challenged by the rise of new technology and there's no doubt that it will have much less prominence in the future" within the IB.[298] Even the use of external markers and 'reflection sessions' within the EE clearly do not offer sufficient protection against plagiarism and other forms of dishonesty.

8. Performance-based assessments

In the context of assessment, 'performances' is a broad concept. Stecher (2010) suggests that it is more helpful to consider what performance assessment is *not;* that is, rather than choosing from predetermined options, performance assessment requires candidates to "either construct or supply an answer, produce a product or perform an activity."[299] As discussed earlier in this report, there has been a deliberative shift by Ofqual towards 'non-exam assessment' (NEA) and away from coursework and other teacher-led assessments, yet within this overall pattern there are several subjects that retain a large proportion of NEA. This chapter will focus on assessments related to 'performing an activity' in two such subjects - music and drama.

Music

Below is a typical specification for A-level music, consisting of three assessment components:

- **A written exam paper** (40 per cent of the marks) – students appraise music during a 2 hour and 30 minute exam, which is then externally marked.
- **A performance** (35 per cent) – this must last a minimum of 10 minutes, as either a solo and/or ensemble as an instrumentalist, or vocalist and/or music production (via technology). The audio of the performance is recorded and then externally marked by exam board assessors.
- **A composition** (25 per cent) – students must produce two compositions (one to a brief, and one of their own choosing) that have a combined duration of a minimum of four and a half minutes. The composition is also externally marked.[300]

There is a considerable research literature documenting the difficulties in assessing musical performances. McPherson and Thompson (1998) highlighted the complex set of interacting factors that affect performance assessments, including that "significant biases often influence the results."[301] For example, the characteristics of the evaluator can "strongly influence" the outcome, including their personality, experience, musical ability, prior training, familiarity with the performer and familiarity with the repertoire.[302] The characteristics of the performer can also be a source of bias in some cases. Elliott (1995) dubbed identical audio tracks over video recordings of white males and females as well as black males and females playing the trumpet and flute, and subsequently asked 88 music education majors to score the performances. Black performers were scored significantly lower than their white peers, and female trumpeters scored lower than female flutists. The research concluded that "prior expectation can influence how even experienced musicians hear and judge musical

performances".[303] However, there are ways to avoid these issues such as providing anonymous audio recordings of performances to limit the opportunity for bias.

Alongside the concerns over bias, McPherson and Thompson noted that "reliability among assessors is sometimes low". [304] They cited the work of Fiske (1978) who presented a set of performances twice to experienced musicians. Assessors were not made aware that the performances were repeated and thus provided two judgements for each one. Correlations between the first and second set of judgements were "alarmingly low", with some assessors showing a negative correlation between their own ratings. This led Fiske to suggest that assessors may have been applying "inconsistent" criteria when judging performances.[305]

Thompson et al. (1998) looked what criteria ('constructs') expert assessors ('adjudicators') used to evaluate music performances.[306] Adjudicators listened to recorded performances of a piece of music by Frédéric Chopin and then devised six constructs to assess the music; five of which were supposed to refer to specific characteristics of the performance (e.g. tempo), whereas one should capture overall quality.[307] There was some overlap between the constructs chosen by the adjudicators, but also some differences.[308] In the second part of the study, adjudicators listened to six different recordings of the same piece of music and assessed each one on a seven-point 'Likert' scale based on their own chosen constructs. A comparison of the various "overall preference" constructs showed only a moderate correlation (0.53) between the six adjudicators,[309] suggesting that experienced musicians may develop their own 'internal' mark scheme even when they are being asked to complete the same task such as provide a holistic mark. Further work by Thompson and Williamon (2003) looked at the correlations between assessors when marking musical performances according to a mark scheme.[310] The assessors' marks showed positive correlations (between 0.332 and 0.651), but this range was described as "rather moderate" across the full set of mark scheme categories.[311]

Researchers have also looked at whether reliability is affected by the approach to marking. Fiske (1977) compared the evaluations of trumpet performances made using a mark scheme that included an 'overall quality' mark which had no specified relationship to the other marking categories. Fiske found there was greater variation between evaluators on the various marking categories than for the 'overall quality' score. This could imply that "evaluators internally weighted the segmented categories differently in arriving at their overall mark" or "interpreted the meaning of the categories differently."[312] In a similar study, Mills (1991) asked assessors to rate a musical performance both holistically and according to a 12-category mark scheme.[313] The study found that there was no direct relationship between the holistic mark and the category marks awarded by assessors. Mills suggested that there may therefore be "no assessment advantage" in categorised mark schemes because they "may not adequately reflect the process of arriving at a holistic, overall mark."[314] Mills also argued that holistic assessment is more "musically credible" than categorised marking as it appears

"closer to the kind of informal quality judgements made in everyday listening."[315] Swanwick (1996) also emphasised the complexity of musical experiences, stating that "such a rich activity cannot be reduced to a single dimension" and warned that "when we conflate several observations we lose a lot of important information on the way."[316]

Despite the challenges associated with judging musical performances, it remains an accepted and widely used method of assessment due to its high validity and credibility. As Thompson and Williamon argue, "performance quality… is meaningless without real-world validity"[317] and that one "may have to accept" that performance assessment is "simply not open to… reliable and consistent scrutiny".[318] In effect, this trade-off between high validity and real-world relevance on the one hand and low reliability on the other hand is accepted within this form of performance assessment because any attempt to increase the reliability could jeopardise the purpose of the assessment itself.

Drama

Below is a typical specification for A-level Drama and Theatre:

- **An open book written exam** (worth 40 per cent of the marks) – students are assessed on their knowledge and understanding of how drama and theatre is developed and performed during a 1 hour 45 minute paper.
- **Creating an original drama** (30 per cent) – students must create and perform their own drama, with the marks split between the performance itself and an assessment of their working notebook documenting the process of creating the drama. These are marked by teachers and moderated by the exam board.
- **Making theatre** (30 per cent) – students must perform and interpret three extracts from different plays as well as keep a 'reflective report'. Only the third extract is formally presented to an audience and is marked by a visiting exam board assessor, as is the reflective report. [319] [320]

The assessment of drama performances presents "unique challenges".[321] As Carroll and Dodds state, "unlike the ability to add or subtract numbers, creativity cannot be taught explicitly, and is also difficult to measure systematically".[322] Jacobs (2016) states that one of the difficulties with assessment in 'aesthetic domains' such as drama is that it "utilises personal responses to stimuli, which can be unfamiliar to those more accustomed to assessment tasks with previously defined answers."[323] While traditional assessments focus on objectivity "whereby assessors are expected to discard their own feelings in favour of strictly set criteria",[324] in this case "a focus on objective judgements is contrary to drama education".[325] This can result in challenges with achieving good reliability because assessors having different

interpretations of what constitutes a 'good' drama performance may reduce the consistency of marks between them. Consequently, Tomlinson (2001) argued for a balance between subjective and objective judgements in performance assessments that provide the most "individually sensitive, accurate and comprehensive evidence."[326]

While subjectivity can evidently cause problems in relation to the reliability of drama assessments, there are ways of overcoming these issues. Baptiste (2008) argued that subjectivity is a "natural aspect" of arts assessment, and that its impact "can be managed through the use of rubrics and the training of assessors."[327] Similarly, Clark (2002) claimed that "well-constructed" assessment criteria can provide a way for students to be evaluated "easily and equitably, without compromising on the divergent qualities of individual creative processes" that arise in drama assessment tasks.[328] However, some researchers have argued against the use of criteria in performance assessments. According to Amabile (1996) and Sternberg (1988), "any products derived from a known formula or pre-determined set of instructions can never be considered creative."[329]

In addition to the risk of subjectivity, the assessment of drama performances can face further obstacles to achieving good reliability. For example, the emotions of students on the day of the assessment can affect their performance. Johnson and Emunah (2009) argued that performance anxiety among actors can cause them to "freeze on stage, alter their behaviour or fail to perform in the manner that they had planned."[330] Among drama students, Jacobs (2014) identified multiple sources of anxiety that can affect their performance including anxiety about their own performance (e.g. how well they remember their lines), anxiety about creative work (e.g. how well the audience will receive the performance) and anxiety about being assessed (e.g. whether they will do well).[331] Jacobs found that students often report feeling a variety of emotions – not just negative ones – including being nervous, happy, worried, pleased, confused, excited and scared. These emotions have the potential to "inspire or inhibit the student's performance, possibly affecting their results."[332]

Reliability and validity challenges can also arise during the marking process. As with other assessment methods discussed in this report, the assessors can be a source of bias. Leach et al. (2000) argued that both conscious and unconscious bias can arise from assessors' own values, preferences and dispositions,[333] leading to lower reliability between assessors and potentially lower validity if assessors are not making fair judgements about a student's attainment. Further problems can arise due to the "fleeting nature of performative assessment."[334] As Jacobs (2014) heard from a drama teacher, "you can see the piece again, but your reaction will be different because you know what to expect… when it's over, the bubble pops, you're out of their performance space, and you have to recall what you saw, or felt."[335] The student's teacher must therefore balance the roles of assessor and audience member when making judgements about the student's performance. What's more, teachers emphasised that video

recordings of performances provided "a different experience to the live performance" and could be "clinical", with emotional responses not always elicited from recordings.[336] The teachers felt this "has implications for using video recordings for assessment purposes."[337] This could be particularly relevant for the validity of drama assessments, as there is a risk that the full experience is not being captured (and thus measured) when performances are recorded.

Although achieving good reliability and validity in the assessment of drama performances is far from straightforward, there are important reasons why this assessment method is used. Speck (1998) suggested that "to criticise subjectivity is to undercut professional judgement itself, and to recommend its replacement by a reliability test is to devalue its importance."[338] Performance assessment in drama also sets out to achieve credibility and reflect real-world settings; to assess drama without a performance element would arguably be an omission of the essence of the subject itself. Moreover, most drama students choose the subject due to "an interest in engaging in performance or creative work" and, as Lovesy (2002) states, they "simply love 'to do'".[339] Similarly, Jacobs (2016) argued that teachers and students "engage in drama and performance to experience the joy of creative expression and artistic creation, to play 'pretend' in a range of roles and to build a more comprehensive understanding of the human experience through an array of lenses."[340] As with assessing musical performances, it seems that drama assessments do not seek to (or claim to) deliver high levels of reliability because the focus remains firmly on real-world relevance and credibility.

9. Recommendations

Just like written exams, the various alternative assessment methods analysed in this report make trade-offs between validity, reliability, real-world applicability, practicality and credibility. Looking across all the alternative forms of assessment, several cross-cutting themes have emerged in terms of their potential role in our high-stakes assessment system:

- **Developing and demonstrating wider skills:** a common criticism of written exams is their focus on recalling knowledge, whereas alternative methods often emphasise other competencies. For example, the EPQ aims to develop research skills, extended writing and presentation skills as well as encouraging students to work independently, and oral assessments give students an opportunity to demonstrate their knowledge in a more practical way while also developing verbal communication skills. What's more, a student's performance on, say, a portfolio or oral assessment may not always be consistent with the same student's performance in a written exam, which suggests that different methods of assessment could be measuring knowledge and understanding in a different way.

- **Reflecting the 'real world':** written exams are normally completed in silence in an artificial environment that does not reflect real-world settings such as the workplace. In contrast, the nature of creative subjects such as art and design, drama and music means that it is necessary to use assessments that prioritise their relationship to real-world contexts (e.g. live performances) rather than prioritising reliability. That said, even though such assessments may have less reliability than written exams, they are still a more credible way to measure a student's attainment in these subjects.

- **Preventing malpractice:** several alternatives to written exams struggle to guarantee that the work a student submits is their own. As highlighted by the historical problems with coursework, malpractice can quickly become widespread and threaten the validity of an assessment in a high-stakes system. Similar concerns over malpractice apply in the present day to the EPQ and the IB extended essay, even more so with the emergence of 'chatbots'. This compares unfavourably with written exams, where the scope for malpractice is greatly reduced by its controlled testing environment.

- **Delivery issues:** coursework-style tasks and extended projects are often hampered by their inconsistent delivery across schools and colleges, as teachers may (inadvertently or otherwise) provide some students with more support and guidance than others, even within the same institution. Numerous attempts have been made in the past to reduce these inconsistencies, but written exams retain a considerable advantage in being able to ensure that all students across the country receive the same questions in a standardised manner during the assessment.

- **The practicality of assessments:** written exams are relatively cheap to deliver and mark, which is beneficial when assessing large numbers of students in a short timeframe. The same cannot be said for coursework and controlled assessment, which were shown to have a significant impact on curriculum time during the academic year as well as creating a greater workload for teachers. Moreover, the options for improving the reliability of some alternatives to written exams can have a positive effect (e.g. having multiple markers for a portfolio), yet the increase in the time and investment required to mark all students' work in this manner is likely to prove undeliverable in practice.

- **The risk of inflated grades:** internally marked assessments tend to produce higher grades than external assessments such as written exams. This increase in grades has often led to a 'bunching' of marks at the top end of the grade distribution for many internal assessments including coursework, controlled assessments and the EPQ. As a result, it becomes harder to differentiate between students and typically leads to grade inflation – most notably during the COVID-19 pandemic. If an assessment does not differentiate between students, this could have significant implications including a qualification losing its value over time or universities and employers introducing their own entrance tests.

- **Disparities between students:** as demonstrated by many academic studies and the pandemic, any assessment marked by a student's own teacher is likely to be less valid and reliable than an external assessment. This is due to the risk of inconsistent marking by teachers coupled with the prospect of bias related to demographic factors such as a student's socio-economic background or a teacher's knowledge of a student's prior grades. External assessors can also be 'biased' to some extent, particularly during more subjective assessments for creative subjects. In most cases, such bias can be largely, if not entirely, overcome by using anonymous marking (e.g. assessing an audio recording of a musical performance, or having a written exam assessed anonymously by an external marker).

- **Different approaches to marking:** while challenges clearly arise when teachers are asked to award scores / grades to students in a high-stakes system, this report has found evidence that teachers are nevertheless able to accurately rank students based on their performances. Reliability may also be improved by drawing on assessors' expertise through asking them to make holistic judgements about a student's performance instead of judging their performance against detailed marking criteria.

By combining these cross-cutting themes, the following pages will describe a package of reforms that is designed to:

(i) build on the strengths of written exams;

(ii) draw on the strengths offered by various alternative assessment methods – which in many cases directly address some of the weaknesses with written exams set out at the start of this report.

RECOMMENDATION 1

To maintain the credibility of the high-stakes assessment system in the final years of secondary education, written examinations should continue to be the main method of assessing students' knowledge and understanding. In contrast, placing a greater emphasis on coursework and other forms of 'teacher assessment' would increase teachers' workload and lead to less reliable grades that may be biased against students from disadvantaged backgrounds.

Any assessment that is used for high-stakes summative purposes must produce trustworthy judgements on the attainment of students. While every method of assessment in mainstream education involves some form of trade-offs, this report has established that the pressures of a high-stakes system reduce the credibility of many alternatives to written exams – particularly those marked by a student's own teacher. The research evidence clearly demonstrates that the scope for malpractice, bias, inconsistent delivery and increased teacher workload – to name just a few challenges – are greatly increased outside of standardised, externally marked written exams. On that basis, written exams should retain their position as the main method of conducting summative assessments in our high-stakes system. Where non-exam assessment (NEA) is necessary, the government should continue to follow the principles set out by Ofqual described in chapter 3 i.e. NEA should only be used where it is the only valid way to assess essential elements of particular subjects (e.g. performances in A-level music), and wherever it is used, NEA should be designed so that the final grades are not easily distorted by the high stakes nature of the assessment.

RECOMMENDATION 2

To broaden the curriculum and develop a wider range of skills than those promoted by written exams, students aged 16-19 taking classroom-based courses should be required to take one additional subject in Year 12 (equivalent to an AS level) that will be examined entirely through an oral assessment.

As noted in chapter 2, verbal communication is among the skills that employers most commonly say is lacking among school and college leavers. The majority of pupils studying academic and applied subjects (e.g. A-levels and BTECs) have limited opportunities to develop these skills as their courses are largely assessed through written exams. Meanwhile, T-level students and apprentices are normally assessed using a mixture of methods such as presentations, vivas, interviews and professional discussions that can enhance their verbal communication skills alongside other attributes. Given their importance for future progression and employment, there is no good reason why these skills should not be developed among all students.

Consequently, this report calls for every student enrolled in classroom-based courses such as A-levels to study one additional subject that will be assessed through an oral assessment instead of a written exam. There are several options as to how this could work:

- **Option 1**: students will study one additional subject relative to current academic and applied programmes (e.g. four A-level subjects rather than three) during their first year of post-16 education. At the end of the first year, they will choose one subject to take as an AS-level and complete an oral assessment on what they have learned on that course, after which they will no longer study that subject. They will then take mostly written exams (depending on the subjects being studied) in their remaining academic and applied subjects at the end of their second year.
- **Option 2**: students will study one additional subject relative to current academic and applied programmes across both years of post-16 education. At the end of the second year, they will complete an oral assessment in one subject of their choice and sit mostly written exams in their other subjects.
- **Option 3**: Previously EDSK has called for the introduction of a 'baccalaureate' to reduce the degree of specialisation in the final years of secondary education. In this model students will be studying five 'academic' subjects or three 'applied' subjects (or a combination of them) in their penultimate year of secondary education and would then choose one subject to be tested through an oral assessment at the end of that year. Their remaining four academic or two applied subjects would be assessed largely through written exams at the end of their final year.

All the above options come with pros and cons, but the overriding goal is clear: students would take an oral assessment on a subject of their choice that would cover the equivalent of an AS-level course.

In terms of how the oral assessments would be delivered, this report suggests drawing on the approach to existing assessments such as GCSE language orals and the German *Abitur*. Although a detailed consultation would need to take place before finalising arrangements for the new oral assessments, it is envisaged that there will be a window in which schools and colleges carry out the oral assessments during the summer term and the assessment itself will involve something akin to the following process:

- The assessment will last for 20 minutes and will consist of two parts: (i) a presentation; and (ii) interview-style questions. The assessment will be carried out by a teacher(s) based at the school or college.
- For the presentation, students will get 30 minutes to prepare immediately before their oral assessment. Students will be offered several options for a presentation topic and will select one topic, at which point they can start preparing their notes without access to additional support materials e.g. textbooks.

- Students will deliver their presentation at the start of the oral assessment. Questions will only be asked by the assessor if the candidate has misunderstood the topic.
- For the interview, the teacher-examiner will ask the student a range of questions that are selected from a question bank provided by the exam board, as is the case for foreign language GCSE and A-level exams. These questions will be based on course material unrelated to the topic that the student has just presented on. Students will not see the interview questions in advance of the assessment.

Every assessment will be audio-recorded by the teacher-examiner. It is worth considering whether an independent observer would be needed during the assessment to ensure that the appropriate processes are being followed, although this would have obvious logistical implications. An even more rigorous approach would be to have a single independent assessor (i.e. a teacher from another school or college) carry out the oral assessments, with schools and colleges being grouped locally to facilitate this process, but again this would be more complicated to deliver than using existing staff.

After every student has completed their oral assessment, the teacher-examiner(s) will make a holistic judgement about every student's performance during the whole assessment and then rank the students being assessed orally in that subject. This ranking will be based on a combination of two factors: (i) the quality of a student's knowledge and understanding of the course material; and (ii) the quality of the student's oral communication skills. In the most popular subjects, this may require several assessors (teachers) to work together to rank a cohort of students. After the students have been ranked for each subject, the audio-recordings and rankings will be sent to the exam board, which will assign scores to the students and award them their final grade. This methodology is explicitly designed to draw on the expertise of teachers (ranking their students) but without putting them in the compromising position of having to choose their own students' grades within a high-stakes system.

RECOMMENDATION 3

To ensure that students taking classroom-based subjects can develop their research and extended writing skills beyond an exam setting, the Extended Project Qualification (EPQ) should be made compulsory. In future, the EPQ will be used as a low-stakes skills development programme and will therefore be ungraded.

Recent debates around the introduction of a 'Baccalaureate' in the final years of secondary education have highlighted the potential role for independent research projects as a core component of a Baccalaureate framework.[341] This is unsurprising, given the evidence showing that the EPQ is often praised for engaging students in their learning by giving them the chance

to work and learn independently. Meanwhile, the IB includes a compulsory independent essay because it is seen as a valuable exercise for students alongside their subject-specific courses and wider experiences such as community service. In addition, T-level students and apprentices complete assessments that are relevant to the workplace and develop a broad range of skills through practical assignments and employer-designed projects. Despite these converging approaches, the benefits of completing an independent project are currently restricted among classroom-based students to those who actively choose to pursue the EPQ alongside their other subjects. Consequently, this report calls for the EPQ to become a compulsory qualification for all students studying classroom-based subjects at Level 3 (equivalent to A-levels). As with the existing EPQ, it will be a standalone course in which students research a topic of their choosing, producing the final output of either a written report or an artefact. As now, students will be largely expected to work independently and would need to choose a topic that is outside the syllabus of the subjects they are studying.

As explored earlier in this report, it is difficult to avoid a 'bunching' of marks at the top end of the distribution with coursework-style assessments, and grade inflation is a common issue with teacher assessment more broadly. Moreover, when students are competing for university places, there may be an even greater incentive to commit malpractice. Given these inevitable pressures from a high-stakes system, it is recommended that the EPQ is no longer graded, with students awarded a simple pass/fail judgement instead. Although some university applicants currently receive lower offers on the condition of achieving a high grade in their EPQ, the proposal to make the EPQ compulsory will mean that completing an extended project is no longer a differentiating factor between students and the grade they achieve will thus no longer a central feature of their application.

It could be argued that treating the EPQ as compulsory rather than optional may affect students' motivation but, as the IB has already shown, there is considerable value in students being required to engage in this form of independent project to broaden their research skills. That a student can select a topic of their choice for this extended project also creates a new source of personal motivation that is generally unattainable within most subject specifications. In addition, the recent reforms to A-level science have shown that not attaching an overall grade to practical tasks does not necessarily result in less commitment from students and may even lead to a greater improvement in their independent learning capabilities over time.

In terms of skills development, one could simply expand the existing EPQ to more students without alteration, there is a case for adapting the qualification so that it is more directive about the precise skills that students are expected to develop as well as ensuring that these skills are demonstrated by all students. In the reformed A-level science specifications, every student is required to complete 12 practical experiments and the specifications set out the skills that students must develop (e.g. devising and investigating testable questions; using

specialist equipment to take measurements). Teachers are also required to keep a record of details such as when each practical activity was undertaken and which students met the criteria. On that basis, the current logbook that EPQ students complete could be enhanced so that it ensures students are consistently engaged in the research process and demonstrate certain competencies (e.g. analytical skills) at specific points rather than waiting until they hand in their final project. This approach would help deliver the objective of a compulsory EPQ being used as a low-stakes skills development programme that prepares young people for Higher Education and employment.

RECOMMENDATION 4

To give schools and colleges the resources they need to expand their 16-19 curriculum to include an additional subject and the EPQ, the 'base rate' of per-student funding (currently £4,642) should be increased by approximately £200 a year to reach £6,000 by 2030.

It will not be possible to deliver the recommendations in this report within the existing funding settlement for schools and colleges. England already has a limited number of funded teaching hours in the 16-19 phase relative to many other countries,[342] meaning that any attempt to expand the courses and qualifications available to classroom-based students without additional funding is unlikely to lead to high-quality provision.

First and foremost, students taking more courses will mean more students being entered for final assessments. For example, introducing a compulsory EPQ would see entries dramatically increase from close to 40,000 now to perhaps over 300,000 students each year. The entry fee per pupil is £62.45,[343] meaning that around £16 million of additional investment would be needed each year to cover these fees. Similarly, an AS/A level oral exam to be conducted by an exam board currently costs £40.[344] Last year, there were around 280,000 students entering A-level exams and a further 120,000 entering Applied General exams (e.g. BTECs),[345] although 80,000 A-level students only studied one or two subjects[346] (suggesting that they combined these subjects with Applied General courses). This produces a rough estimate of 320,000 students being entered for an additional oral exam as a result of this report's recommendations – costing around £13 million a year.

However, these additional exam entry fees would be overshadowed by the cost of delivering an extra subject for all students on classroom-based courses relative to the status quo. Since 2010, funding for 16-19 education has been dramatically tightened. From 2010-11 to 2013-14, 16-19 spending per student was, at best, static in cash terms. A new funding formula was introduced in 2013-14, and between then and 2019-20 the national 'base rate' per student was frozen at £4,000 in cash terms, eroding its real value by 9 per cent.[347] In 2019-20 the base rate was eventually increased but only by around £200 per student,[348] costing the government

about £200 million.[349] The base rate is now set to rise to £4,642 per student in the 2023/24 academic year,[350] although this is equivalent to an increase of just 2.2 per cent – well below the current level of inflation. Regardless of these welcome increases to per-student funding over the last few years, college spending per student in 2024–25 will still be around 5 per cent below 2010–11 levels, while school sixth-form spending per student will be 22 per cent below 2010–11 levels.[351] In other words, the funding settlement for 16-19 education remains woefully inadequate.

To address this longstanding concern, EDSK recently called on the government to raise the per-student base rate for 16 to 19-year-olds to £6,000 to ensure that schools and colleges can provide the quality of teaching and learning that young people need and deserve. To reach this funding goal, the government will have to increase the base rate by approximately £200 per student every year for the next seven years; representing around £1.4 billion in additional spending each year by 2030. Even this increase may end up being relatively modest if inflation remains high in the coming years, but it is nevertheless the least that a current or future government could do to broaden the curriculum for 16 to 19-year-olds and provide the necessary resources for the new qualifications and assessments that this report has proposed.

Conclusion

> *"It is a reasonable thing that a school should be judged in part by the ability of its pupils to reach certain intellectual standards prescribed by independent authorities and competently tested by impartial examinations. But the results of the latter are not in themselves sufficient evidence that a school provides the course of intellectual training which affords the most lasting benefit to the pupils, or gives to them the best preparation for the tasks of later life."* [352]

Over 100 years since its publication, this quote from the landmark 'Acland Report' still resonates today. Written exams occupy a pivotal role in our education system and the evidence presented throughout this report shows that there are several important reasons for this; namely, their independent administration, standardised nature, impartial marking and low cost of delivery. Nevertheless, it must also be acknowledged that written exams have limitations, such as the lack of breadth in the skills they promote and the artificial conditions in which pen-and-paper assessments are normally conducted. This report has demonstrated that several alternatives to written exams offer a legitimate (and, in many cases, preferable) way for students to develop intellectual skills such as independent research and verbal communication, which are often highly prized by employers and thus likely to prove beneficial later in life.

That said, the debate over the benefits and drawbacks of written exams cannot avoid the fact that many alternative methods of assessment struggle to withstand the pressures generated by a high-stakes system. Historically, one of the biggest challenges faced by any coursework or essay-style assessments has been ensuring that the work a student submits is indeed their own. This challenge has arguably now become insurmountable in the age of ChatGPT and other 'chatbots', which can produce on-demand material that aims to mimic human dialogue and other creative outputs (both written and non-written). On that basis, replacing written exams with most of the alternative assessment methods explored in this report would almost certainly deliver final grades that are less accurate and trustworthy than those produced by exams while also adding significant new workload burdens onto teachers.

Some of the alternative methods discussed in this report undoubtedly remain useful tools to support teaching and learning through low-stakes formative assessments. Even so, the prospect of greater inaccuracies and inconsistencies in grading within high-stakes summative assessments is surely intolerable when, rightly or wrongly, those same grades form part of the annual competition among students for a finite number of places at many universities and

employers. What's more, taxpayers have every right to expect that a publicly funded assessment system is delivering fair judgements on students, even more so when the government uses examination results to help identify which schools and colleges are performing well and which are potentially struggling. That is not to say that the grades awarded for written exams are perfect in this regard, but the alternatives to exams explored in this report generally fare worse than exams in terms of their ability to differentiate between students in a precise and consistent manner.

Changing the national assessment system will inevitably involve some degree of disruption for students, teachers, leaders and parents. On that basis, any proposed reforms should be expected to demonstrate that they would add value beyond the existing assessments while also protecting the interests of learners and taxpayers. The recommendations in this report thus seek to build on the strengths of written exams while drawing on the strengths of other forms of assessment that can also operate within a high-stakes system: oral assessments and low-stakes independent research projects. Through this approach, students would engage with a broader suite of courses and assessments that give them the opportunity to develop additional skills and capabilities over and above those fostered by written exams.

Admittedly, such progress will only be possible if a current or future government is willing to invest in our secondary education system to ensure that every institution has sufficient resources to deliver the proposals outlined in this report. This would require undoing the funding cuts seen in 16-19 education over the past decade or so, which would be a laudable objective in its own right. Should this new investment materialise, the combination of written exams, oral exams and independent research projects proposed in this report will create a solid foundation for our high-stakes assessment system in schools and colleges for many years to come.

References

[1] Wikiquote, 'Albert Einstein', Webpage, 2022.

[2] FE News, 'GCSE, AS and A Level Exams 2021 Will Go Ahead with 3 Week Delay', Webpage, 12 October 2020; Department for Education, '"Exams Are the Best and Fairest Way for Young People to Show What They Know and Can Do" – The Education Secretary on the Importance of Exams', 29 November 2020; Sophia Sleigh, 'Schools Minister: Exams May Not Be Back to Normal in 2022 for Children Disrupted by Pandemic', *Evening Standard*, 25 February 2021; William Stewart, 'GCSEs and A Levels: Exams "Are Fairest", Says Ofqual', *Times Educational Supplement*, 6 January 2021.

[3] Thomas Kellaghan and Vincent Greaney, *Public Examinations Examined* (Washington DC: World Bank Group, 2020), 44.

[4] Ben Wilbrink, 'Assessment in Historical Perspective', *Studies in Educationul Evaluation* 23, no. 1 (1997): 42.

[5] Wikipedia, 'Imperial Examination', 2021.

[6] Wilbrink, 'Assessment in Historical Perspective', 32.

[7] Ibid., 35.

[8] Ibid., 36.

[9] Stephen Dobson, 'The Life and Death of the Viva', in *Assessing the Viva in Higher Education: The Enabling Power of Assessment*, vol. 6 (Springer, Cham, 2018), 151.

[10] A.H. Dyke Acland, *Report of the Consultative Committee on Examinations in Secondary Schools* (London: Her Majesty's Stationery Office, 1911), 6.

[11] Dobson, 'The Life and Death of the Viva', 152.

[12] Ibid., 151.

[13] Acland, *Report of the Consultative Committee on Examinations in Secondary Schools*, 26.

[14] Wikipedia, 'Mathematical Tripos', 2021.

[15] Wilbrink, 'Assessment in Historical Perspective', 39.

[16] Acland, *Report of the Consultative Committee on Examinations in Secondary Schools*, 20.

[17] Ibid., 7–8.

[18] Ibid., 13.

[19] Ibid., 15.

[20] Ibid., 17.

[21] Ibid., 103.

[22] Ibid.

[23] Derek Gillard, 'Education in England - Chapter 7: 1900-1923', Webpage, 2018, 7.

[24] Acland, *Report of the Consultative Committee on Examinations in Secondary Schools*, 114.

[25] Ibid., 123.

[26] Derek Gillard, 'Education in England - Chapter 10: 1945-1951', Webpage, 2018.

[27] Wikipedia, 'Certificate of Secondary Education', Webpage, 2020.

[28] Qualifications and Curriculum Authority, 'The Story of the General Certificate of Secondary Education (GCSE)', Webpage, 2007.

[29] Wikipedia, 'Raising of School Leaving Age in England and Wales', Webpage, 2020.

[30] Acland, *Report of the Consultative Committee on Examinations in Secondary Schools*, 102.

[31] Ibid.

[32] Ibid.

[33] Department for Education, 'Employer Skills Survey 2019: England Results', Webpage, 2020.

[34] Institute for Apprenticeships and Technical Education, 'Developing an Occupational Standard', we, 2021.

[35] Sally Weale, 'Stress and Serious Anxiety: How the New GCSE Is Affecting Mental Health', *The Guardian*, 17 May 2018.

[36] House of Commons Children, Schools and Families Committee, *Testing and Assessment: Third Report of Session 2007–08 - Volume I*, HC169-I (London: Her Majesty's Stationery Office, 2008), 49.

[37] Ofsted, 'HMCI's Commentary: Recent Primary and Secondary Curriculum Research', Webpage, 11 October 2017.

[38] Ofqual, *Research and Analysis: Marking Consistency Metrics - An Update* (Coventry: Ofqual, 2018).

[39] Will Hazell, 'Will GCSEs Be Scrapped? Why England's Exam System Could Face Major Overhaul after Covid Results Fiasco', *The i Paper*, 18 October 2021.

[40] Council of Skills Advisers, *Learning and Skills for Economic Recovery, Social Cohesion and a More Equal Britain* (London: Labour Party, 2022), 38.

[41] Harris MacLeod, 'Public Tepid on "Gove Levels"', *YouGov*, 27 September 2012.

[42] Mark Enser, 'GCSEs and A Levels 2021: Can Exams Ever Be Fair?', *Times Educational Supplement*, 14 October 2020.

[43] Barnaby Lenon, 'Rethinking Assessment by Barnaby Lenon', *The University of Buckingham*, 28 September 2020.

[44] Dennis Opposs, 'Whatever Happened to School-Based Assessment in England's GCSEs and A Levels?', *Perspectives in Education* 34, no. 4 (2016): 54.

[45] Ibid.

[46] Ibid.

[47] Ibid.

[48] Ibid.

[49] Ibid., 55.

[50] Ofqual, *Review of Controlled Assessment in GCSEs*, 2013, 5.

[51] Nick Davies, 'Fiddling the Figures to Get the Right Results', *The Guardian*, 11 July 2000.

[52] Ofqual, *Review of Controlled Assessment in GCSEs*, 5.

[53] Ian Colwill, *Improving GCSE: Internal and Controlled Assessment* (Qualifications and Curriculum Authority, 2007), 3.

[54] Ofqual, *Review of Controlled Assessment in GCSEs*, 6.

[55] Ibid., 7.

[56] Ipsos MORI, *Evaluation of the Introduction of Controlled Assessment* (Ofqual, 2011), 2.

[57] Ibid., 3.

[58] Jeevan Vasagar, 'GCSE Results Cause Outcry over "unfair" Marking', *The Guardian*, 23 August 2012.

[59] Ofqual, *Review of Controlled Assessment in GCSEs*, 7.

[60] Ibid., 10.

[61] Ofqual, *GCSE Reform Consultation*, 2013, 44.

[62] Ofqual, *Review of Controlled Assessment in GCSEs*, 11.

[63] Ofqual, *GCSE Reform Consultation*, 20.

[64] Ofqual, *Review of Controlled Assessment in GCSEs*, 12.

[65] Ibid., 13.

[66] Ofqual, *GCSE Reform Consultation*, 20.

[67] Ofqual, *Review of Controlled Assessment in GCSEs*, 13.

[68] Ibid., 14.

[69] Ibid., 15.

[70] Ibid.

[71] Ofqual, *GCSE Reform Consultation*, 21.

[72] Opposs, 'Whatever Happened to School-Based Assessment in England's GCSEs and A Levels?', 59.

[73] Ofqual, 'Summary of Changes to GCSEs from 2015', Webpage, 31 March 2017.

[74] Opposs, 'Whatever Happened to School-Based Assessment in England's GCSEs and A Levels?', 60.

[75] Ofqual, *Consultation on the Assessment of Practical Work in GCSE Science*, 2014, 5.

[76] Ofqual, 'Summary of Changes to AS and A Levels from 2015', Webpage, 5 June 2018.

[77] Ofqual, *Confirmed Assessment Arrangements for Reformed AS and A Level Qualifications*, 2014, 6.

[78] William Stewart, 'Practical Science to Be Removed from A Levels Due to Fears of Cheating and Over-Marking', *Times Educational Supplement*, 25 October 2013.

[79] Opposs, 'Whatever Happened to School-Based Assessment in England's GCSEs and A Levels?', 59.

[80] Stuart Cadwallader, *The Impact of Qualification Reform on the Practical Skills of A Level Science Students* (Ofqual, 2019), 6.

[81] Ibid.

[82] Nicola Woolcock, 'ChatGPT Marks End of Homework at Alleyn's School', *The Times*, 29 January 2023.

[83] Karen MacGregor, 'Sciences Po Bans ChatGPT amid HE Quality, Integrity Fears', *University World News*, 3 February 2023.

[84] Samantha Booth, 'Ofqual Chief Would Use Exam Conditions for Coursework amid ChatGPT Fears', *Schools Week*, 10 March 2023.

[85] Gavin Williamson, *Impact of Covid-19 on Summer Exams*, HCWS176, 2020.

[86] Williamson, *Impact of Covid-19 on Summer Exams*.

[87] Ibid.

[88] Amelia Hill and Caroline Davies, 'A-Level Results Day 2020 Live: 39.1% of Pupils' Grades in England Downgraded - as It Happened', *The Guardian*, 13 August 2020.

[89] Sally Weale and Heather Stewart, 'A-Level and GCSE Results in England to Be Based on Teacher Assessment in U-Turn', *The Guardian*, 17 August 2020.

[90] Sean Coughlan, 'A-Levels and GCSEs: Boris Johnson Blames "mutant Algorithm" for Exam Fiasco', *BBC News Online*, 26 August 2020.

[91] Ofqual, *Results Table for GCSE, AS and A Level Results in England, 2020*, 2020.

[92] Ibid.

[93] Ofqual, 'How Qualifications Will Be Awarded in 2021', Webpage, 25 February 2021.

[94] Ibid.

[95] Ofqual, 'Infographics for GCSE Results, 2021', Webpage, 2021.

[96] Martin Robinson, 'MP Says Lack of Learning during Pandemic Has Been a "national Disaster" for UK's Poorer Pupils as Traditional State Secondaries Get Half as Many A-Level A-Grades as Private Schools', *Mail Online*, 10 August 2021.

[97] Matilda Martin, 'GCSEs 2021 Teacher Grades Were Seen as "Unreliable"', *Times Educational Supplement*, 27 October 2022.

[98] Ofqual, *A Level Outcomes in England*, 2021.

[99] Ofqual, 'Summer 2021 Results Analysis and Quality Assurance - A Level and GCSE', Webpage, 2021.

[100] Ibid., 21.

[101] Matilda Martin, 'Exam Boards Investigate Private Schools over Grade Malpractice', *Times Educational Supplement*, 12 October 2022.

[102] Ming Wei Lee, *Summer 2021 Student-Level Equalities Analysis - GCSE and A Level* (Ofqual, 2021).

[103] Ibid.

[104] The Sutton Trust, 'The Sutton Trust Comment on Results Day 2021', Webpage, 2021.

[105] Catherine Lough and Amy Gibbons, 'GCSE and A Level Results 2021: What Did Teachers Learn?', *Times Educational Supplement*, 2021.

[106] John M Malouff and Einar B Thorsteinsson, 'Bias in Grading: A Meta-Analysis of Experimental Research Findings', *Australian Journal of Education* 60, no. 3 (2013): 245–56.

[107] Lord Bew, *Independent Review of Key Stage 2 Testing, Assessment and Accountability* (London: Department for Education, 2011), 49.

[108] Tammy Campbell, 'Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment', *Journal of Social Policy* 44, no. 3 (2015): 517–47.

[109] Alberto Alesina et al., *Revealing Stereotypes: Evidence from Immigrants in Schools* (Cambridge, MA: National Bureau of Economic Research, 2018), 3.

[110] Ineke Pit-ten Cate and Sabine Glock, 'Teachers' Attitudes towards Students with High- and Low-Educated Parents', *Social Psychology of Education* 21, no. 3 (July 2018).

[111] Lewis Doyle, Matthew J. Easterbrook, and Peter R. Harris, 'Roles of Socioeconomic Status, Ethnicity and Teacher Beliefs in Academic Grading', *British Journal of Educational Psychology* 93, no. 1 (23 August 2022): 91–112.

[112] Paul Newton, 'Bias in Teacher Assessment Results', *Ofqual*, 17 May 2021.

[113] Ibid.

[114] Bew, *Independent Review of Key Stage 2 Testing, Assessment and Accountability*, 50.

[115] Ibid., 51.

[116] Susan M Brookhart, 'The Use of Teacher Judgement for Summative Assessment in the USA', *Assessment in Education: Principles, Policy & Practice* 20, no. 1 (2013): 69–90.

[117] House of Commons Education Committee, *Primary Assessment: Eleventh Report of Session 2016-17*, HC-682 (London: Her Majesty's Stationery Office, 2017), 10.

[118] Ibid.

[119] Helen Johnson, 'Teacher-Assessed Grades: "Exhausted" Teachers Hit out at Increased Workload Due to GCSE and A Level Replacement', *National World*, 4 May 2021.

[120] Gordon Joughin, *A Short Guide to Oral Assessment* (Leeds Metropolitan University, 2010), 1.

[121] Ibid., 5.

[122] Ibid., 10.

[123] Ibid.

[124] Gordon Joughin, 'Dimensions of Oral Assessment', *Assessment and Evaluation in Higher Education* 23, no. 4 (1998): 371.

[125] Joughin, *A Short Guide to Oral Assessment*, 1.

[126] Joughin, 'Dimensions of Oral Assessment', 372.

[127] Ibid., 372–73.

[128] Joughin, *A Short Guide to Oral Assessment*, 12.

[129] Kultusministerkonferenz, *Guidelines for the Regulations Concerning the Achievement of the Allgemeine Hochschulreife (General Higher Education Entrance Qualification) at German Schools Abroad: German International Abitur* (Kultusministerkonferenz, 2015), 26.

[130] Wikipedia, 'Abitur', Webpage, 2023.

[131] AQA, *GCSE French* (AQA, 2016), 8.

[132] Wynne Harlen, *A Systematic Review of the Evidence of Reliability and Validity of Assessment by Teachers Used for Summative Purposes* (Evidence for Policy and Practice Information and Co-ordinating Centre, 2004), 47.

[133] Ibid., 48.

[134] Ibid.

[135] Ibid.

[136] Brian Richards and Francine Chambers, 'Reliability and Validity in the GCSE Oral Examination', *Language Learning Journal*, no. 14 (1996): 31.

[137] Ibid.

[138] Ibid.

[139] Ibid., 29.

[140] Ibid.

[141] Ayesha Ahmed, Alastair Pollitt, and Leslie Rose, *Assessing Thinking and Understanding: Can Oral Assessment Provide a Clearer Perspective?* (Cambridge Assessment, 1999), 2.

[142] Ibid., 9.

[143] Ibid., 10.

[144] Ibid.

[145] Ibid.

[146] Mark Huxham, Fiona Campbell, and Jenny Westwood, 'Oral versus Written Assessments: A Test of Student Performance and Attitudes', *Assessment and Evaluation in Higher Education* 00, no. 0 (2010): 9.

[147] Bernie Carter and Karen Whittaker, 'Examining the British PhD Viva: Opening New Doors or Scarring for Life?', *Contemporary Nurse: A Journal for the Australian Nursing Profession*, July 2009, 8.

[148] Ibid., 9–10.

[149] Ibid., 4–5.

[150] Brian Poole, 'Examining the Doctoral Viva: Perspectives from a Sample of UK Academics', *London Review of Education* 13, no. 3 (2015): 92.

[151] Ibid.

[152] Ibid., 95–97.

[153] Sharmilla Torke et al., 'The Impact of Viva-Voce Examination on Students' Performance in Theory Component of the Final Summative Examination in Physiology', *Journal of Physiology and Pathophysiology* 1, no. 1 (April 2010): 10.

[154] Ibid., 12.

[155] Ayaz Khurram Mallick and Ayyub Patel, 'Comparison of Structured Viva Examination and Traditional Viva Examination as a Tool of Assessment in Biochemistry for Medical Students', *European Journal of Molecular and Clinical Medicine* 7, no. 6 (2020): 1786.

[156] R.-A Knight, L Dipper, and M Cruice, 'The Use of Video in Addressing Anxiety Prior to Viva Voce Exams' 44, no. 6 (2013).

[157] Jane Mellanby and Anna Zimdars, 'Trait Anxiety and Final Degree Performance at the University of Oxford', *High Educ* 61 (2011): 357.

[158] Mallick and Patel, 'Comparison of Structured Viva Examination and Traditional Viva Examination as a Tool of Assessment in Biochemistry for Medical Students', 1787.

[159] L Nunnik et al., 'A Prospective Comparison between Written Examination and Either Simulation-Based or Oral Viva Examination of Intensive Care Trainees' Procedural Skills', *Anaesth Intensive Care* 38 (2010): 876.

[160] Ibid.

[161] Ibid., 881.

[162] Torke et al., 'The Impact of Viva-Voce Examination on Students' Performance in Theory Component of the Final Summative Examination in Physiology', 12.

[163] Mallick and Patel, 'Comparison of Structured Viva Examination and Traditional Viva Examination as a Tool of Assessment in Biochemistry for Medical Students', 1788.

[164] Ibid., 1785.

[165] June C Yang and Douglas W Laube, 'Improvement of Reliability of an Oral Examination By a Structured Evaluation Instrument', *Journal of Medical Education* 58 (1983): 864.

[166] Val Wass et al., 'Achieving Acceptable Reliability in Oral Examinations: An Analysis of the Royal College of General Practitioners Membership Examination's Oral Component', *Medical Education* 37 (2003): 130.

[167] Richard Wakeford, Lesley Southgate, and Val Wass, 'Improving Oral Examinations: Selecting, Training, and Monitoring Examiners for the MRCGP', *BMJ* 311 (1995): 935.

[168] Margery H. Davis and Gominda G. Ponnamperuma, 'Portfolio Assessment', *Journal of Veterinary Medical Education* 32, no. 3 (2005): 279.

[169] David Baume, *A Briefing on Assessment of Portfolios* (Learning and Teaching Support Network Generic Centre, 2001), 7–9.

[170] AQA, 'GCSE Design and Technology: Scheme of Assessment', Webpage, 2023.

[171] Ibid.

[172] AQA, 'GCSE Design and Technology: Non-Exam Assessment Administration', Webpage, 2023.

[173] Staci Leon and Maurice Elias, 'A Comparison of Portfolio, Performance, and Traditional Assessments in Middle School', *Research in Middle Level Education Quarterly* 21, no. 2 (1998): 25.

[174] Ibid.

[175] Sharon Buckley et al., *The Educational Effects of Portfolios on Undergraduate Student Learning: Best Evidence Medical Educational (BEME) Systematic Review No. 11* (Medical Teacher, 2009), 7.

[176] Daniel Koretz, 'Large-Scale Portfolio Assessments in the US: Evidence Pertaining to the Quality of Measurement', *Assessment in Education* 5, no. 3 (1998): 315.

[177] Ibid., 316.

[178] Ibid.

[179] Ibid., 319.

[180] Ibid., 320.

[181] Ibid., 321.

[182] Ibid., 318.

[183] Ibid., 322.

[184] Ibid., 323.

[185] Ibid.

[186] Ibid., 324.

[187] Edward W. Wolfe, 'A Report on the Reliability of a Large-Scale Portfolio Assessment for Language Arts, Mathematics and Science' (Annual Meeting of the National Council for Measurement in Education, New York, NY, 1996), 16.

[188] Ibid., 17.

[189] Koretz, 'Large-Scale Portfolio Assessments in the US: Evidence Pertaining to the Quality of Measurement', 4.

[190] David Baume and Mantz Yorke, *Validity and Reliability in the Evaluation of Portfolios for the Accreditation of Teachers in Higher Education* (AAHE Assessment Forum, 2000), 4.

[191] Ibid.

[192] Carol M. Myford and Robert J. Mislevy, *Monitoring and Improving a Portfolio Assessment System* (National Center for Research on Evaluation, Standards, and Student Testing: University of California, Los Angeles, n.d.), 13.

[193] Ibid., 21–26.

[194] Ibid., 76.

[195] Koretz, 'Large-Scale Portfolio Assessments in the US: Evidence Pertaining to the Quality of Measurement', 327.

[196] Ibid.

[197] Ibid.

[198] Ibid.

[199] Ibid., 329.

[200] Ibid.

[201] Ibid., 331.

[202] Lewis Elton and Brenda Johnstone, *Assessment in Universities: A Critical Review of Research* (Learning and Teaching Support Network Generic Centre, 2002), 56.

[203] Ibid.

[204] Koretz, 'Large-Scale Portfolio Assessments in the US: Evidence Pertaining to the Quality of Measurement', 331–32.

[205] Gordon Stanley et al., *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development* (Oxford University Centre for Educational Assessment, 2009), 70.

[206] Koretz, 'Large-Scale Portfolio Assessments in the US: Evidence Pertaining to the Quality of Measurement', 332.

[207] Ibid.

[208] Elton and Johnstone, *Assessment in Universities: A Critical Review of Research*, 37.

[209] Chris Roberts, David I. Newble, and Alan J. O'Rourke, 'Portfolio-Based Assessments in Medical Education: Are They Valid and Reliable for Summative Purposes?', *Medical Education* 36, no. 10 (November 2002): 899.

[210] Margery H. Davis and Gominda G. Ponnamperuma, 'Examiner Perceptions of a Portfolio Assessment Process', *Medical Teacher* 32, no. 5 (2010): 211.

[211] Buckley et al., *The Educational Effects of Portfolios on Undergraduate Student Learning: Best Evidence Medical Educational (BEME) Systematic Review No. 11*, 9.

[212] Ibid., 7.

[213] Baume and Yorke, *Validity and Reliability in the Evaluation of Portfolios for the Accreditation of Teachers in Higher Education*, 3.

[214] Ibid.

[215] Ibid., 7.

[216] Ibid., 9.

[217] Ibid.

[218] Elton and Johnstone, *Assessment in Universities: A Critical Review of Research*, 57.

[219] Claire Tochel and Alex Haig, 'The Effectiveness of Portfolios for Post-Graduate Assessment and Education: BEME Guide No 12', *Medical Teacher*, May 2009, 11–12.

[220] Ibid., 11.

[221] NRM Michels et al., 'Portfolio Assessment during Medical Internships: How to Obtain a Reliable and Feasible Assessment Procedure?', *Education for Health* 22, no. 3 (2009): 1.

[222] Tochel and Haig, 'The Effectiveness of Portfolios for Post-Graduate Assessment and Education: BEME Guide No 12', 12.

[223] Ibid.

[224] Ibid., 13.

[225] Ibid.

[226] Claire Stocks and Chris Trevitt, *Signifying Authenticity: How Valid Is a Portfolio Approach to Assessment?* (Oxford Learning Institute, University of Oxford, 2010), 17.

[227] Tochel and Haig, 'The Effectiveness of Portfolios for Post-Graduate Assessment and Education: BEME Guide No 12', 13.

[228] Stocks and Trevitt, *Signifying Authenticity: How Valid Is a Portfolio Approach to Assessment?*, 5.

[229] Ibid., 7–8.

[230] Tochel and Haig, 'The Effectiveness of Portfolios for Post-Graduate Assessment and Education: BEME Guide No 12', 12.

[231] Ibid.

[232] Ibid.

[233] C. P. M. Van der Vleuten et al., 'A Model for Programmatic Assessment Fit for Purpose', *Medical Teacher*, no. 34 (2012): 206.

[234] Ibid., 207.

[235] Elton and Johnstone, *Assessment in Universities: A Critical Review of Research*, 57.

[236] Tochel and Haig, 'The Effectiveness of Portfolios for Post-Graduate Assessment and Education: BEME Guide No 12', 57–58.

[237] Ibid., 15.

[238] Roberts, Newble, and O'Rourke, 'Portfolio-Based Assessments in Medical Education: Are They Valid and Reliable for Summative Purposes?', 899.

[239] Buckley et al., *The Educational Effects of Portfolios on Undergraduate Student Learning: Best Evidence Medical Educational (BEME) Systematic Review No. 11*, 25.

[240] Roberts, Newble, and O'Rourke, 'Portfolio-Based Assessments in Medical Education: Are They Valid and Reliable for Summative Purposes?', 899.

[241] Department for Education and Skills, *14-19 Education and Skills* (Her Majesty's Stationery Office, 2005), 63.

[242] Ben Jones and Charlotte Stephenson, 'The EPQ and Academic Performance'.

[243] AQA, 'Level 3 Extended Project Qualification: Introduction', Webpage, 2023.

[244] UCAS, 'Extended Project Qualification (EPQ)', Webpage, 2019.

[245] AQA, *Level 3 Extended Project Qualification: Specification* (Manchester: AQA, 2021), 3.

[246] Ibid., 5.

[247] Ibid., 13.

[248] Tim Gill, *Uptake and Results in the Extended Project Qualification 2008-2015* (Cambridge: Cambridge Assessment, 2016), 2.

[249] Joint Council for Qualifications, *Provisional Level 3 Extended Project Results - June 2022* (Joint Council for Qualifications, 2022).

[250] Gill, *Uptake and Results in the Extended Project Qualification 2008-2015*, 5.

[251] Ibid., 4.

[252] Tim Gill, *An Analysis of the Effect of Taking the EPQ on Performance in Other Level 3 Qualifications* (Cambridge: Cambridge Assessment, 2016), 2.

[253] Ibid., 6.

[254] Ben Jones, *Does the Extended Project Qualification Enhance Students' GCE A-Level Performance?* (Manchester: AQA, 2016), 6.

[255] Ibid., 8.

[256] Phoebe Surridge, Caroline Lau, and Yasmine El Masri, *Does the Extended Project Qualification Enhance Students' A-Level Results?* (Manchester: AQA, 2021), 2.

[257] Jones, *Does the Extended Project Qualification Enhance Students' GCE A-Level Performance?*, 3.

[258] Tim Gill, *Are Students Who Take the Extended Project Qualification Better Prepared for Higher Education?* (Cambridge: Cambridge Assessment, 2022), 9.

[259] Ibid., 17.

[260] Ibid., 52.

[261] Tim Gill and Carmen Vidal Rodeiro, *Predictive Validity of Level 3 Qualifications* (Cambridge: Cambridge Assessment, 2014), 28.

[262] Charlotte Stephenson and Tina Isaacs, 'The Role of the Extended Project Qualification in Developing Self-Regulated Learners: Exploring Students' and Teachers' Experiences', *The Curriculum Journal*, 2019, 7.

[263] Ibid., 16.

[264] Ibid., 16–17.

[265] Ibid., 17.

[266] Ibid.

[267] Qingping He and Beth Black, *An Investigation of the Variability in Grade Outcomes in Extended Project Qualification* (Ofqual, 2018), 4.

[268] Ibid.

[269] Ibid., 5.

[270] Ibid., 32.

[271] This data is combined from two sources: (i) Joint Council for Qualifications, *Provisional Level 3 Extended Project Results - June 2022*; (ii) Joint Council for Qualifications, *Applied GCE, AEA, Extended Project Results: Summer 2020* (JCQ, 2020), 6.

[272] This data is combined from two sources: (i) Joint Council for Qualifications, *Provisional Level 3 Extended Project Results - June 2022*; (ii) Ofqual, 'A Level Outcomes in England', Webpage, 2022.

[273] Ofqual, 'A Level Outcomes in England', 2022.

[274] John Roberts, 'UK Students Outperform Global Average in International Baccalaureate', *Times Educational Supplement*, 5 July 2022.

[275] International Baccalaureate, 'Programmes: Diploma Curriculum', Webpage, 2022.

[276] International Baccalaureate, 'Understanding DP Assessment', Webpage, 2022.

[277] International Baccalaureate, 'What Is the Extended Essay', Webpage, 2022.

[278] International School Parent and Barbara Macario, 'The IB Extended Essay Explained', Webpage, 2023.

[279] International Baccalaureate, 'What Is the Extended Essay'.

[280] Ibid.

[281] Ibid.

[282] Mark W. Aulls, David Lemay, and Sandra Peláez, *Research Summary: Exploring the Learning Benefits and Outcomes of the Extended Essay in Preparing Students for University in Canada (Two Phases)* (International Baccalaureate, 2013), 4.

[283] Ibid.

[284] Mary Lee Taylor and Marion Porath, 'Reflections on the International Baccalaureate Program: Graduates' Perspectives', *Journal of Advanced Academics* 17, no. 3 (2006): 149.

[285] Karen Kurotsuchi Inkelas et al., *Exploring the Benefits of the International Baccalaureate Extended Essay for University Studies at the University of Virginia* (Centre for Advanced Study of Teaching and Learning in Higher Education, 2012), 4.

[286] Ibid., 30.

[287] Ibid., 32.

[288] Ibid., 30.

[289] David Wray, *Student Perceptions of the Value of the International Baccalaureate Extended Essay in Preparing for University Studies: Final Report* (University of Warwick, 2013), 25.

[290] Ibid., 27.

[291] Ibid., 29.

[292] Ibid.

[293] Ibid., 5.

[294] Inkelas et al., *Exploring the Benefits of the International Baccalaureate Extended Essay for University Studies at the University of Virginia*, 37.

[295] Ibid., 35.

[296] Selim Algar, 'ChatGPT Cheating Scandal Erupts inside Elite Program at Florida High School', *New York Post*, 16 February 2023.

[297] Louisa Clarence-Smith, 'Schools Could Get Official Chatbot Guidance to Stop Pupils Cheating', *The Daily Telegraph*, 30 December 2022.

[298] Nicola Woolcock, 'International Baccalaureate Lets Pupils Use ChatGPT to Write Essays', *The Times*, 27 February 2023.

[299] Brian Stecher, *Performance Assessment in an Era of Standards-Based Educational Accountability* (Stanford, CA: Stanford Center for Opportunity Policy in Education, 2010), 2.

[300] AQA, 'A-Level Music: Scheme of Assessment', Webpage, 2023.

[301] Gary E. McPherson and William F. Thompson, 'Assessing Music Performance: Issues and Influences', *Research Studies in Music Education* 10, no. 12 (June 1998): 12.

[302] Ibid., 15.

[303] Ibid.

[304] Ibid., 12.

[305] Ibid., 17.

[306] Sam Thompson and Aaron Williamon, 'Evaluating Evaluation: Musical Performance Assessment as a Research Tool', *Music Perception* 21, no. 1 (September 2003): 26.

[307] Ibid.

[308] Ibid.

[309] Ibid.

[310] Ibid., 21.

[311] Ibid., 29.

[312] Ibid., 26.

[313] Ibid.

[314] Ibid.

[315] Ibid.

[316] McPherson and Thompson, 'Assessing Music Performance: Issues and Influences', 19.

[317] Thompson and Williamon, 'Evaluating Evaluation: Musical Performance Assessment as a Research Tool', 38.

[318] Ibid.

[319] AQA, 'A-Level Drama and Theatre: Specification at a Glance', Webpage, 2023.

[320] AQA, 'A-Level Drama and Theatre: Scheme of Assessment', Webpage, 2023.

[321] Rachael Jacobs, 'Challenges of Drama Performance Assessment', *Drama Research: International Journal of Drama in Education* 7, no. 1 (April 2016): 2.

[322] Prerna Carroll and Emma Dodds, 'Taking Risks and Being Creative: Assessment in Drama and Theatre', *Research Matters: A Cambridge Assessment Publication*, 2016, 23.

[323] Jacobs, 'Challenges of Drama Performance Assessment', 4.

[324] Ibid.

[325] Ibid.

[326] Ibid., 5.

[327] Lennise Baptiste, 'Managing Subjectivity in Arts Assessment', in *Reconceptualising the Agenda for Education in the Caribbean* (School of Education, 2008), 508.

[328] Rachael Jacobs, 'Drama Performance Assessment in Senior Secondary Schools: A Study of Six Australian Schools' (University of Western Sydney School of Education, 2014), 36–37.

[329] Ibid., 35.

[330] Ibid., 158.

[331] Ibid.

[332] Ibid., 157.

[333] Ibid., 34.

[334] Ibid., 168.

[335] Ibid., 169.

[336] Ibid., 171.

[337] Ibid.

[338] Baptiste, 'Managing Subjectivity in Arts Assessment', 506.

[339] Jacobs, 'Challenges of Drama Performance Assessment', 4.

[340] Ibid., 13.

[341] Tom Sherrington, 'Towards A National Baccalaureate for England: Building Confidence in an Alternative to the Current Examination System', Webpage, 2022.

[342] Ann Hodgson and Ken Spours, *Tuition Time in Upper Secondary Education (16-19): Comparing Six National Education Systems* (London: UCL Institute of Education, 2016).

[343] AQA, *Entry Fees and Other Charges Summer 2023* (AQA, 2023), 23.

[344] Pearson Edexcel, *Pearson Edexcel Qualifications: Fees for UK Centres* (London: Pearson Education Limited, 2023), 9.

[345] Department for Education, 'Academic Year 2021/22: A Level and Other 16 to 18 Results', Webpage, 2023.

[346] Ofqual, 'Infographics for A Level Results, 2022 (Accessible)', Webpage, 18 August 2022.

[347] Luke Sibieta and Imran Tahir, *Latest Trends in Further Education and Sixth Form Spending in England* (Institute for Fiscal Studies, 2022).

[348] David Foster and House of Commons Library, '16-19 Education Funding in England since 2010', Webpage, 19 February 2020.

[349] HM Treasury et al., 'Chancellor Announces £400 Million Investment for 16-19 Year Olds' Education', Press Release, 31 August 2019.

[350] Jason Noble, '16-19 Base Rate to Rise by Just 2.2% from August 2023', *FE Week*, 9 January 2023.

[351] Institute for Fiscal Studies, 'Further Education and Skills', Webpage, 2023.

[352] A.H. Dyke Acland, *Report of the Consultative Committee on Examinations in Secondary Schools* (London: Her Majesty's Stationery Office, 1911), 137.