Assessment in mathematics in England

An MEI discussion paper

1 CONTENTS

1.1 BACKGROUND

- The appropriateness of students sitting high-stakes external examinations at age 16 when they will continue in education for another two years has been questioned.
- The cancellation of examinations in summer 2020 and 2021 due to the Covid19 pandemic has highlighted the dependence of the English assessment system on end of course, timed written examinations. Other jurisdictions make use of some teacher assessment.

1.2 THE PURPOSE OF THIS PAPER

Whether or not a fundamental change occurs in the assessment system across all subjects, MEI wishes to promote discussion about assessment in mathematics.

This discussion paper raises some of the issues which need to be considered when assessing mathematics attainment within a national summative assessment system. It does not claim to be an exhaustive analysis and it does not attempt to make any recommendations. The aim is to encourage informed discussion of the advantages and disadvantages of the current system in England and how it might be improved. Current assessment practices in England are considered, along with previous practices and some from other countries.

Formative assessment is a vital part of teaching and learning but how this is carried out is not the focus of this paper. However, the possibility of using teachers' assessment of their students as a contribution to national summative assessment is considered.

2 WHAT IS THE PURPOSE OF ASSESSMENT?

2.1 THE PURPOSES OF ASSESSMENT

Assessment can have many purposes. The form of assessment should be aligned to its intended purposes. Where assessment has more than one purpose, as is usual, different forms may be more suitable for each of the different purposes. Some of the main purposes of assessment (current and past) are listed below, in no particular order.

- To give feedback to students, celebrating their successful learning and encouraging them to move on to the next stage of their education
- To motivate students to work hard
- To give feedback to teachers about student learning and inform future teaching
- To indicate whether students are suitable for progression to another course or employment
- To identify the highest achieving students
- To provide a qualification which can be used by students for a variety of purposes
- To determine whether students have met a national standard for their learning
- To compare the effectiveness of different teachers or schools, for example to compare education systems internationally; to evaluate national programs and policy; to support decisions related to teachers' pay and progression.

2.2 THE DISCONTINUITY AT AGE 16

A striking and far-reaching feature of the education system in England is the change which young people experience at age 16. This applies as much in mathematics as in the rest of the curriculum.

The mathematics curriculum up to age 16 has the following features.

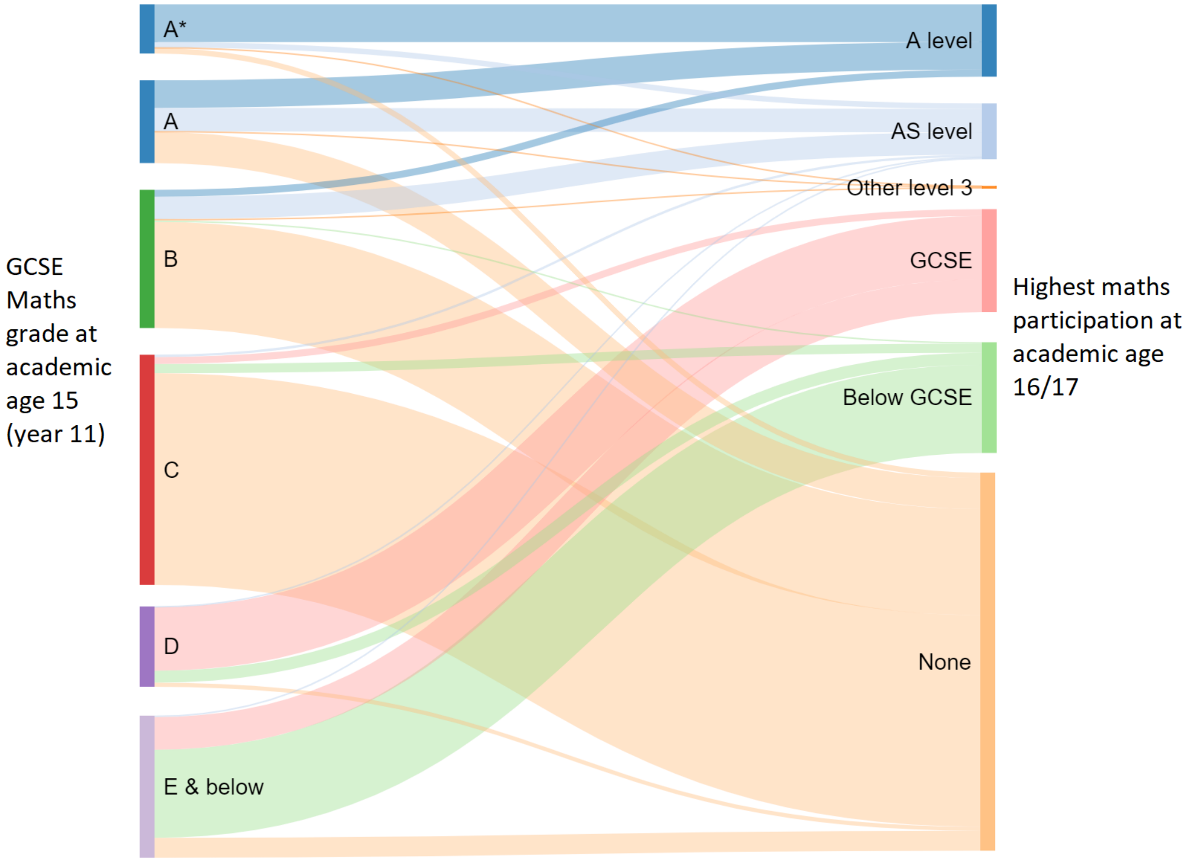
- Learning mathematics is compulsory until age 16.
- There is a common curriculum for all learners until age 16.
- The complete curriculum is designed to ensure that all who complete it successfully are ready for further study of mathematics post-16, including calculus.
- A subset of the complete curriculum is designed to ensure that all those who complete it successfully have sufficient knowledge, skills and understanding in mathematics to apply basic mathematics in future study, work and life.
- Delivery in secondary schools is almost always as a separate subject.
- Delivery in primary schools is increasingly as a separate subject.
- Maths is also taught as needed in other subjects. In some cases, the maths needed is not taught in the common maths curriculum (eg perspective in art); in other cases, the maths may be taught later in the common maths curriculum (eg correlation in geography).
- Maths is externally assessed for all at ages 11 and 16.
- At age 11 outcomes are about achieving an expected standard. Not reaching the expected standard may result in a loss of confidence.
- At age 16 almost all young people are assessed by sitting GCSE Mathematics; this might be at either Foundation or Higher Tier.
- At age 16 the outcomes are assessed against a common scale of grades. These grades are used for different purposes; achieving lower than grade 4 seriously limits pathways for progression for young people in education and employment. Although grades 1 to 3 at GCSE are pass grades, they are often not seen as such. The importance for both students and schools of achieving at least grade 4 means that there is a risk of secondary education focusing on training students for examinations rather than on teaching the intended curriculum.

From 16 to 19, young people might not be working towards a mathematics qualification, but most young people are learning and using maths in some way in their education, whether on an academic or technical pathway.

The Sankey diagram below analyses post-16 maths participation for young people aged 19 in 2016, when GCSEs were graded A* to G. These data[1] were originally produced for the Smith review[2] and count participation in terms of mathematics qualifications; the widths of the bars are proportional to numbers of students.

[1] Post-16 maths participation in 2015 to 2016, Ad-hoc notice, July 2017, DfE

[2] Professor Sir Adrian Smith's review of post-16 mathematics education for 16-to 18-year-olds in England



Post-16 maths participation for different GCSE maths grades

The diagram does not show the variety of the maths being learned or used by those in the 'None' category.

- Some students will be studying A levels which rely on mathematics or quantitative skills but not A level Mathematics. Since the data were collected for Figure 1, Core Maths qualifications have become available for such students. There were around 12,000 entries to Core Maths qualifications in 2020 and numbers are gradually increasing. However, this still leaves tens of thousands of students using maths in their A levels but not pursuing a maths qualification.
- Some students will be following technical or vocational courses, or apprenticeships. Many of these will be using or learning maths but not pursuing a maths qualification.

Some consequences of the discontinuity at age 16 are as follows.

- The GCSE assessment of the curriculum to age 16 is required to meet several purposes, acting as preparation and a gatekeeper for all students for several maths and non-maths qualifications and pathways. It is also meant to signify that a student has sufficient knowledge, skills and understanding in mathematics to apply basic mathematics in future study, work and life. Other qualifications by contrast, for example A level Mathematics, have a smaller cohort and more focused set of purposes.
- There is considerable pressure on schools and teachers to ensure students attain at least grade 4 at age 16 and this can distort the implemented curriculum.
- Those students who do not attain grade 4 at age 16 will have reduced opportunities for future education and employment[3] but many have expressed concern that the resitting of GCSE Mathematics by large numbers of post-16 students is not fit for purpose, has a negative effect on attitudes to maths education and damages public perception of

[3] Review of Vocational Education – The Wolf Report, Alison Wolf, 2011

mathematics. Almost 180,000 students resat GCSE Mathematics in summer 2019 but only 22.3% attained a level 2 pass (grade 4 or above).[4]
- A small number of other qualifications, for example Additional Mathematics, have been developed to challenge students who are ready for more than Higher Tier GCSE Mathematics at age 16. In summer 2020, there were 32,301 candidates in total for AQA Level 2 Certificate in Further Mathematics and OCR FSMQ Additional Mathematics.
- Many learners on post-16 courses would benefit from following a mathematics qualification alongside their other qualifications but do not choose to do so. This may be due to poor advice and/or a lack of confidence due to a perception that they have not done well in maths pre-16 or to the lack of a suitable option in their post-16 institution. It is important to examine the curriculum and the assessment pre-16 to determine whether they are significant factors in producing any lack of confidence among young people in their mathematical understanding.

2.3 Assessment at 16: could we drop GCSE Mathematics?

There is a growing national conversation about 'dropping GCSEs'. There may be a case for treating mathematics differently from most other subjects because of the importance of mathematics in supporting most post-16 pathways and employment.

One possible outcome of a serious review of assessment at 16 might be that a national assessment system remains for (say) English and mathematics, with alternative arrangements to mark attainment in other subjects. This might allow a system to be devised which works well for assessing mathematics, rather than the current one-size-fits-all approach. The tiering arrangement and exam-only approach of the current GCSE qualifications have come about from consideration of the whole GCSE system; perhaps there is an opportunity to decide what works best for mathematics.

A poor outcome of any such review of assessment at 16 would be to design more appropriate ways of marking attainment in other subjects but to keep the current GCSE arrangements for English and mathematics. This emphasises the importance of the mathematics education community revisiting the issues around assessment in mathematics – hence this paper.

The rest of this paper makes no assumptions about what happens to the national system. It considers the advantages and disadvantages of the current assessments, particularly at GCSE, and how they might be improved.

2.4 Curriculum and assessment: what do we want to assess?

The UNESCO International Bureau of Education glossary explains curriculum coherence as follows.

A characteristic of curriculum indicating the extent to which the curriculum aims and content, as well as textbooks, teaching methods, and assessment are all aligned and reinforce one another. Some research findings suggest that a high level of curriculum coherence is associated with high performing systems. (Adapted from: Oates 2010)[5]

---

[4] Joint Council for Qualifications (2019). *GCSE (Full Course) Results Summer 2019*

[5] http://www.ibe.unesco.org/en/glossary-curriculum-terminology

Consideration of what would be appropriate assessment for mathematics cannot be separated from consideration of what would be an appropriate mathematics curriculum. As noted above, the complete pre-16 curriculum, tackled by students who enter for Higher Tier GCSE Mathematics, was historically designed in line with the needs of students going on to study A level Mathematics. This explains the emphasis on algebra. It could be argued that many students would benefit from a greater focus on quantitative skills. This discussion is beyond the scope of this paper.

This feature of the pre-16 curriculum, that it is designed to be preparation for A level Mathematics, results in a top-down approach to curriculum design. Students with a good understanding of GCSE Mathematics at Higher tier are well prepared for progression to AS and A level Mathematics but these students are not the majority; students with grade 7 or above are the ones who are most likely to consider taking Mathematics at A level but this is only about 16% of candidates[6]. A similar situation was considered almost 40 years ago in the Cockcroft report[7].

> ***In our view, very many pupils in secondary schools are at present being required to follow mathematics syllabuses whose content is too great and which are not suited to their level of attainment****. Efforts to introduce pupils to as much of the examination syllabus as possible result in attempts to cover the ground too fast for understanding to develop. The result is that very many pupils neither develop a confident approach to their use of mathematics nor achieve mastery of those parts of the syllabus which should be within their capability.*

The Cockcroft report recommended that curriculum development in mathematics should have a bottom-up approach, starting with consideration of the mathematics which all students need to know and then adding content for students who gain an understanding of the foundational content. The original three tiers of GCSE Mathematics were an attempt to put this recommendation into practice but the importance of gaining grade C led to a reduction to two tiers.

More recently, use of the mastery approach to mathematics teaching is focusing again on the development of mathematical understanding.[8]

> *Mastering maths means students of all ages acquiring a deep, long-term, secure and adaptable understanding of the subject. The phrase 'teaching for mastery' describes the elements of classroom practice and school organisation that combine to give students the best chances of mastering maths. Achieving mastery means acquiring a solid enough understanding of the maths that's been taught to enable students to move on to more advanced material.*
>
> *Teaching for mastery is currently more widespread in primary schools than in secondary schools, but every year hundreds of secondary teachers and schools are taking part in professional development to enable them to deliver teaching for mastery to their students. Secondary schools are also increasingly finding that Year 6s are coming up from primary school having experienced teaching for mastery in maths.*

[6] JCQ GCSE results summer 2019
[7] Mathematics Counts, HMSO, 1982
[8] https://www.ncetm.org.uk/professional-development/school-leaders/

The mastery materials available through NCETM[9] include materials to enable teachers to assess the mathematical understanding of their students. This kind of formative assessment could be incorporated into the overall summative assessment of students. The use of teacher assessment is considered more fully later in this paper.

3 HOW SHOULD ASSESSMENT BE USED TO MAKE JUDGEMENTS?

3.1 VALIDITY AND RELIABILITY

Two key concepts relating to assessment are validity and reliability. Both are concerned with the assessment as a measure of something. A valid assessment measures the right thing. A reliable assessment measures accurately. These concepts are explained by Ofqual as follows.

> *Reliability is a measure of the consistency of the results of qualifications and assessments. A reliable qualification would mean that a student would receive the same result if they took a different version of the exam, took the test on a different day, or if a different examiner had marked the paper.*[10]
>
> *The validity of a particular qualification is the degree to which it is possible to measure whatever that qualification needs to measure by implementing its assessment procedure.*[11]

This raises the question of what we want to measure by using a particular qualification. Perhaps this is whether students are ready for the next stage of learning or whether they have learned enough mathematics to be ready for work or perhaps we want a measure of how well they have learned the intended curriculum.

Ofqual note the interplay between curriculum, teaching, learning and assessment.

> *Qualifications do not operate in a vacuum, independently of other educational concerns. The four pillars of education – curriculum, teaching, learning and assessment – need to operate in synergy with each other. It is especially important that assessment design decisions – however sensible from a validity perspective – do not impact unduly upon curriculum, teaching or learning in such a way as to threaten the acquisition of the very learning outcomes that the qualification is supposed to certify.*[12]

3.2 MODELS OF ASSESSMENT

This section introduces two contrasting models of measurement in assessment: the quality model and the difficulty model. In a quality model, students are judged on how well they have performed a task; in a difficulty model, they are judged on how difficult a task they can do. Ahmed and Pollitt describe the distinction as follows[13].

[9] https://www.ncetm.org.uk/teaching-for-mastery/mastery-materials/
[10] https://www.gov.uk/government/collections/ofquals-reliability-research
[11] An approach to understanding validation arguments, Ofqual, October 2017
[12] An approach to understanding validation arguments, Ofqual, October 2017
[13] Ahmed, Ayesha and Pollitt, Alastair(2010), The Support Model for interactive assessment

To make the distinction between these two models clearer, consider two sporting paradigms that represent them well: ice dance (free skating) and high jump. Competitive ice dancing uses a pure quality model. All ice rinks are equally flat and roughly the same size, shape and temperature; in other words, the task is pretty much the same in every ice dance performance. The skater is expected to go out and perform in a way that impresses the judges as much as possible. In contrast, a high jump competition is a clear example of the difficulty model, consisting of a series of tasks of ever increasing difficulty which continues until everybody has failed. In the ice dance we focus on judging the responses, while in high jump we focus on counting successes on the tasks.

3.3 CRITERION REFERENCED V NORM REFERENCED ASSESSMENT

This section introduces some language to describe an important feature of any assessment system, whether based solely on examinations or not. The UK driving test is criterion referenced; the candidate is assessed as to whether they can perform the tasks deemed necessary for being a safe and competent driver. An athletics race is norm referenced; rank order determines the award of the medals.

Whether mathematics assessments are criterion referenced or norm referenced should perhaps depend on the purpose of the assessment. An assessment to test whether candidates have attained a national standard of mastery of ‘essential maths’ might be criterion referenced. An assessment to determine the best 200 students for admission to a university course might be norm referenced.

Glaser[14] outlined the difference between criterion referenced assessment and norm referenced assessment in 1963.

The scores obtained from an achievement test provide primarily two kinds of information. One is the degree to which the student has attained criterion performance, for example, whether he can satisfactorily prepare an experimental report, or solve certain kinds of word problems in arithmetic. The second kind of information that an achievement test score provides is the relative ordering of individuals with respect to their test performance for example, whether Student A can solve his problems more quickly than Student B. The principal difference between these two kinds of information lies in the standard used as a reference. What I shall call criterion-referenced measures depend upon an absolute standard of quality, while what I term norm-referenced measures depend upon a relative standard.

Glaser made the distinction between criterion referenced assessments, which tell us what a student can do, and norm referenced assessments, which tell us how a student compares to other students. The terms are often used in a narrower sense nowadays. Criterion referenced assessment is often taken to mean that students have shown that they are able to achieve **all** the relevant learning outcomes, as shown in the following extract from the Ofqual blog[15].

Criterion referencing involves measuring a student’s performance against pre-determined criteria, or learning outcomes - written descriptions of what students should know and be able to do. For a qualification to be criterion referenced, the

[14] GLASER, R. (1963) Instructional technology and the measurement of learning outcomes: some questions. Reproduced in Educational Measurement: Issues and Practice, 13 (4), 6-8
[15] https://ofqual.blog.gov.uk/2017/03/17/mythbusting-3-common-misconceptions/

criteria would have to be very clear so that all those involved in assessing students had a common understanding of what was required.

In GCSE the closest thing we have to such criteria are grade descriptors. Here's an example from the legacy GCSE maths grade descriptor for grade C:

"Learners use a range of mathematical techniques, terminology, diagrams and symbols consistently, appropriately and accurately. Learners are able to use different representations effectively and they recognise some equivalent representations for example, numerical, graphical and algebraic representations of linear functions; percentages, fractions and decimals. Their numerical skills are sound and they use a calculator accurately. They apply ideas of proportionality to numerical problems and use geometric properties of angles, lines and shapes."

In a criterion-referenced qualification, each of these statements would need to be met. So a student who could not correctly answer the questions that required them to "use geometric properties of angles, lines and shapes" would not achieve a grade C, no matter how many marks they scored on other questions.

Criterion referencing is in effect a series of mini hurdles that students have to get over. If they fall at one of them, they miss out on that qualification, or that grade.

The same Ofqual blog[16] shows a different understanding of norm-referencing to that outlined by Glaser[17].

It's also worth saying that GCSEs and A levels are not norm-referenced either. If they were, we'd see fixed proportions of each grade in each specification, regardless of the ability of the cohort or the ability profile of any one exam board.

The blog[18] goes on to explain that GCSEs are compensatory qualifications.

That's why GCSEs have never been criterion referenced. GCSEs are 'compensatory' qualifications so better performance in one area can compensate for poorer performance in another.

Wiliam[19] distinguishes between criterion referencing, where the criteria collectively and exhaustively define levels of attainment and situations where the criteria are exemplary rather than exhaustive. He refers to the latter situation as construct-referenced.

The touchstone for distinguishing between criterion- and construct- referenced assessments is the relationship between the written descriptions and the domains. Where written statements collectively define the level of performance required (or more precisely where they define the justifiable inferences), then the assessment is criterion-referenced. However, where such statements merely exemplify the kinds of inferences that are warranted, then the assessment is, to an extent at least, construct-referenced.

---

[16] https://ofqual.blog.gov.uk/2017/03/17/mythbusting-3-common-misconceptions/
[17] GLASER, R. (1963) Instructional technology and the measurement of learning outcomes: some questions. Reproduced in Educational Measurement: Issues and Practice, 13 (4), 6-8
[18] https://ofqual.blog.gov.uk/2017/03/17/mythbusting-3-common-misconceptions/
[19] Assessing authentic tasks: alternatives to mark-schemes, Dylan Wiliam, Nordic Studies in Mathematics Education Vol 2 , No I, 48-68

Criterion referenced assessment is common in vocational qualifications where students must demonstrate particular competences before they can pass the qualification. It is also used in A level science for the practical assessment.

> *In order to be awarded a Pass a Learner must, by the end of the practical science assessment, consistently and routinely meet the criteria in respect of each competency listed below. A Learner may demonstrate the competencies in any practical activity undertaken as part of that assessment throughout the course of study.*[20]

For A level science, and typically in vocational assessments, learners have more than one opportunity to demonstrate the relevant competence. Assessments are overseen/monitored by awarding organisations.

Ensuring that students have demonstrated competence in each one of a set of criteria is one way of implementing criterion referenced assessment; this becomes more difficult to implement when the set of criteria is the whole mathematics curriculum.

Agreeing a set of criteria which corresponds to a particular level of attainment is a non-trivial task but levels of attainment were set out in early iterations of the National Curriculum for Mathematics in England. What would it mean to attempt to test whether a student had met such a level of attainment in a timed written examination? It is not possible to assess the whole curriculum in a timed examination at the end of the course, so a sample of the subject matter is assessed. This raises questions of whether all samples are of equal difficulty, do different students perform differently depending on the sampling and, for a given sample, is one set of possible attainment tasks equally difficult to another? Everyone makes mistakes so even a student who understands everything at a particular level of attainment cannot be expected to score 100% on a sample of questions assessing the relevant content; how much of the assessment does a student need to get right in order to attain the level? These questions raise the same kinds of issues as are common when setting grade boundaries for compensatory qualifications.

3.4 Maintaining standards from year to year

Having established that GCSEs (and A levels) are compensatory qualifications, we now consider what it means that standards are maintained from year to year and whether this is important.

In GCSE and A level examinations, and in other qualifications such as Core Maths, the marks from a set of exam papers put the candidates in a rank order; statistical information is used to help set grade boundaries to ensure that standards are consistent from year to year, i.e. to ensure that any change in the distribution of grades awarded reflects a change in the level of mathematics learning achieved by the cohort concerned. The requirement for consistency in standards is not necessarily regarded as important in other countries. Wolf[21], writing in the year 2000 about the United States describes the situation as follows.

---

[20] GCE Subject Level Conditions and Requirements for Science (Biology, Chemistry, Physics) and Certificate Requirements, May 2016, Ofqual

[21] Wolf, A. (2000) 'A comparative perspective on educational standards', Proceedings of the British Academy, 102, 1-8.

In spite of the explosion of test use, American students are only very rarely in a position where results on these tests have high stakes consequences. A very large part of the marking, grading, and certification that takes place in American education is completely separate from the testing industry's activities. High school diplomas are given essentially on the basis of grades awarded by class teachers, who may take some notice of test results (and often do not: Firestone 1998). Although many states are now introducing state-wide tests which must be passed in order to graduate, these are minimum-competency tests, and as such, low-stakes for most candidates. For the majority, their Grade Point Average is far more important than simple acquisition of a high school diploma; what it records is the average of entirely teacher-given grades. College degrees are gained, again, on the basis of teacher-awarded grades, obtained on a course by course (module by module) basis. In all these contexts, the English observer is struck by the teacher's total autonomy, and by the absence of moderation, quality checks, or even the most minimal form of double marking.

Writing about examinations for university entrance in China, Wolf[22] notes that year on year comparability of standards is not important; the key thing is that candidates in a given year are ranked fairly. The same examination is taken by about 2.5 million students at the same time.

Wolf[23] concludes that the concern with consistency of standards of assessment in England is unusual and that norm referencing, in the sense of comparing candidates, is more important elsewhere.

England—and indeed the UK—is extremely unusual in its overt claim that 'standards' are being maintained from year to year in some absolute sense, and in the primacy of criterion-related concerns over norm-referencing practice. Other countries may make the implicit assumption that 'standards' are being held constant, in the sense of some measure which yields the same quantity year on year; but they certainly do not make models of performance or notions of benchmarking the centrepiece of their item-writing, examining and moderating procedures in the way the UK examination boards have done for many decades. Conversely, other countries tend to be much more overtly focused on the process of differentiation and selection, through a more or less explicit norm-referencing approach. It is commonplace to argue that certificates such as A levels have the dual function of certification and selection, and that the latter function is dominant. However, compared to most countries, what is striking about the UK is that we tend to spend most of our effort on procedures and concerns which are more relevant to certification than to selection pure and simple.

Having an understanding of whether standards are changing over time is important if we want a particular grade in a qualification to mean the same over time but it is also important in understanding whether national standards of attainment are rising or not. The National Reference Test (NRT)[24] has been developed in mathematics and English Language to provide evidence of year-to-year and long-term change in the performance of year 11 students in England. The first NRT took place in 2017, following trialling in 2015 and 2016. A representative sample of students is tested each year using the same test.

---

[22] Wolf, A. (2000) 'A comparative perspective on educational standards', Proceedings of the British Academy, 102, 1-8.
[23] Wolf, A. (2000) 'A comparative perspective on educational standards', Proceedings of the British Academy, 102, 1-8.
[24] National Reference Test Information, Ofqual, NFER, 2019

4 HOW SHOULD ASSESSMENTS BE ORGANISED?

4.1 LINEAR, MODULAR AND WHEN-READY ASSESSMENT: TIMING OF ASSESSMENTS

The discontinuity in education at 16 and the current exam-only approach has resulted in nearly all young people taking a high stakes GCSE examination in mathematics at age 16. This section considers whether taking all the assessments at the end of Year 11 is the best approach.

Considering the different purposes for GCSE might lead to alternative models to achieve different purposes. One purpose of GCSE Mathematics is to demonstrate that a young person has mastered the knowledge, skills and understanding in mathematics necessary to apply basic mathematics in future study, work and life – henceforth called ‘essential maths’. Currently, achieving grade 4 in GCSE Mathematics is taken as a proxy for this. To obtain grade 4, students study (at least) the content of Foundation Tier mathematics, which includes some trigonometry, solving some quadratic equations by factorising and solving simultaneous linear equations; these would not usually be regarded as part of ‘essential maths’. The maths and English condition of funding for 16-19 courses for students who have not attained grade 4 at GCSE mathematics and English means that tens of thousands of young people are studying these topics post-16 rather than concentrating on the mathematics they need to support their other qualifications.

Some have argued that a when-ready assessment of a better-designed ‘essential maths’ qualification, separate from GCSE Mathematics, would be more appropriate. This assessment could be criterion referenced rather than norm referenced. Such an assessment might reduce both the pressure of a multi-purpose GCSE examination and some of the stigma associated with post-16 GCSE resit. A suitable assessment would have the potential to allow all students to have their achievements recognised by age 18.

It could be argued that a better-designed tiering system for GCSE could deal partially with this concern, with the content of one of the tiers being more closely-aligned to ‘essential maths’.

Another issue around the timing of assessments is whether examinations should be linear or modular. The most recent experience of high-stakes modular assessments in England is the AS/A level system which existed prior to the 2017 reforms, so evidence relating to that is considered next.

End of course assessment is currently the preferred mode for GCSE and A level in England but cancellation of examinations in summer 2020 showed up one of the drawbacks of relying solely on an end of course assessment.

Following the 2017 reforms, which decoupled AS from A level, there has been a dramatic decline[25] in the numbers of students taking AS examinations in England. The reduction in numbers combined with the need to write separate exam papers for AS has led to pressure on exam board finances, resulting in AS levels being discontinued in some subjects.

There are advantages to having AS as a part of A level; the information it gives about student progress is useful for teachers, students and other stakeholders such as universities.

[25] From 2016 to 2020, AS Mathematics entries and AS entries overall dropped by 94% in England.

Higton et al[26] found that the modular system was perceived to have advantages such as feedback for students and personalisation of the course but also disadvantages such as encouraging learning for the next module rather than learning the course as a whole. They identified concerns about resitting examinations as follows.

- ***Grade inflation*** *– an increase in the number of students achieving higher grades (see Coe, 2007 for a statistical analysis of grade inflation) was felt by some teachers and higher education sector staff partly to be a consequence of resitting to improve grades. Grade inflation was believed to make it harder to differentiate between high-achieving students. This was of particular concern to HEIs that select students as opposed to those that recruit them;*
- ***Volume of examinations*** *– resitting can lead to a very congested personal examination timetable for A level students, especially in the important second year. Resits increase the number of examinations taken by a student and were felt to limit the amount of time available for study towards the new parts of the syllabus covered and assessed for the first time. This was of particular concern to A level teachers. Interviewees also expressed concerns that, for some students, resits had a detrimental effect on their ability to think more holistically about their subject because they were concentrating too much on passing exams.*

Steps can be taken to address these concerns within a modular system. For example, the WJEC A level in Mathematics for teaching from 2017 limits the number of possible resits of AS level to one and the AS is worth 40% of the whole A level. This contrasts with the former specification where resits were limited only by the availability of examinations and AS counted for 50% of the whole A level.

In 2019, MEI conducted a Nuffield Foundation funded project to investigate what kind of curriculum would be most appropriate for GCSE resit students.[27] As part of the project, qualifications, similar in standard to GCSE, from the past and from other countries were reviewed. We noticed that all qualifications reviewed, with the exception of current GCSEs, were either modular or had a similar lower level qualification which could be used as a stepping stone. This allowed students to experience some success and then make further progress but the current GCSEs do not have a natural stepping stone; this hinders the less confident resit students. When modular GCSEs were available, they were taken by almost half of the candidates.

4.2 Structure of assessment in GCSE Mathematics: models for restricted-grade examinations

Throughout the existence of GCSE Mathematics, the examinations have been tiered with assessment at each tier having a restricted range of grades. It would seem likely that any examination system aimed at nearly all students at 16 would require differentiation by task, and so different combinations of papers would have different grade sets available.

[26] Higton, J., Noble, J., Pope, S., Boal, N., Ginnis, S., Donaldson, R. and Greevy, H. (2012). Fit for Purpose? The view of the higher education sector, teachers and employers on the suitability of A levels. Ofqual.

[27] A new mathematics GCSE curriculum for post-16 resit students, Davies et al, MEI, 2020

Opposs et al[28] discuss four ways of assessing GCSE Mathematics students by examination, providing them with opportunities to demonstrate what they know, understand and can do.

- *Common papers: All pupils take the same papers which are targeted at the full range of grades available, regardless of their ability. Pupils can therefore access the full range of grades with equal opportunities.*
- *Core plus extension paper: All pupils take the core paper which is targeted at the lower grades and more able pupils take the extension paper which is targeted at the higher grades. Only pupils who take the extension paper can access the higher grades.*
- *Tiered papers (the current tiered GCSE model): Pupils can enter for one of the tiers which are targeted at different ranges of available grades, and different tier pupils take different papers with access to the corresponding range of grades.*
- *The adjacent levels model (the Scottish Standard Grade examinations): Pupils can enter for two adjacent levels (or tiers) and papers differ for different levels. Each level is targeted at two adjacent grades and there are no overlapping grades between levels.*

Common examination papers are suitable for subjects where differentiation is by outcome and so would not be suitable for assessing GCSE Mathematics as a whole, where differentiation is by task. We currently have tiered examination papers, so it is worth considering the possible advantages of the other two examination models.

Opposs et al[29] discuss the advantages and disadvantages of all four models and summarise as follows.

> ***Core plus extension paper***
> *The main advantages of this model include minimising the floor and ceiling effects associated with the current GCSE tiering model, avoiding the problem of between-tier comparability, allowing adequate discrimination at the top of the ability range, and minimising any potential negative curriculum backwash effects. The core can be targeted at pupils of middle and low levels of attainment, while the extension paper at the top end pupils.*
>
> ***The adjacent levels model***
> *The adjacent levels model could be viewed as one variant of the core plus extension paper model, but it extends both upwards for high achieving pupils and downwards for low achieving pupils. Assuming there are three levels, the middle level will be taken by most pupils (except the highest and lowest achieving pupils) and therefore acts as the core. The top level would be an extension for higher achieving pupils, while the bottom level for low achieving pupils. A lower level acts as a 'safety net'.*
>
> *The main advantages of the adjacent levels model include minimising the floor and ceiling effects associated with the current GCSE tiering approach, no need to ensure the comparability of overlapping grades, and minimising any potential negative curriculum backwash effects. This model, or a variant of it, would seem particularly suitable for subjects where the differences in demand and content and skills across*

[28] Assessing pupils at the age of 16 in England – what is the best approach for effective examinations? Dennis Opposs, Qingping He, Matt Glanville and Angela Deavall
[29] Assessing pupils at the age of 16 in England – what is the best approach for effective examinations? Dennis Opposs, Qingping He, Matt Glanville and Angela Deavall

the full attainment range of the pupils are large and differentiation between pupils through outcomes will not be effective.

Advantages and disadvantages of the adjacent levels model are considered further by Bramley[30].

4.3 How can we ensure that GCSE Mathematics is at an appropriate level of challenge?

Decisions made about tiering or using one of the alternative approaches affect the number of questions which a candidate can access on a paper and hence the number of marks they might need to attain a particular grade. Mathematics papers typically have easier questions at the beginning with the questions getting gradually harder. The more grades a paper has available, the fewer questions are accessible to candidates getting the lowest grades available. This can bring into question the validity of the examination – are the candidates at the lowest grades available on a paper genuinely showing what they understand and can do? Candidates not being able to access most of the questions on a paper can have a severe impact on the confidence and enjoyment of the young people preparing for and sitting the paper, and this will influence their choice about continuing with the subject.

The grade boundaries from the Pearson/Edexcel GCSE Mathematics examinations in June 2019 are shown below as an example; this is the latest set of grade boundaries from the GCSE specification with the highest number of candidates. The first table shows marks out of 240 and the second shows the boundaries as a percentage, rounded to the nearest whole number. As well as showing some low thresholds for some grades, it also highlights that some candidates are awarded the same grade despite having had very different experiences. Similar grade boundaries are seen in other years and for other GCSE Mathematics awarding bodies' specifications; the low grade boundaries are part of the design of the current GCSE.

Raw marks	Max mark	9	8	7	6	5	4	3	2	1	U
Foundation	240					184	149	111	73	36	0
Higher	240	198	167	137	108	80	52	38			0

Percentage (rounded to integer)	Max mark	9	8	7	6	5	4	3	2	1	U
Foundation	100					77	62	46	30	15	0
Higher	100	83	70	57	45	33	22	16			0

June 2019 Pearson/Edexcel GCSE Mathematics grade boundaries[31]

Grade boundaries change from year to year to reflect the difficulty of the papers. Nevertheless, the way the papers are designed means that low boundaries for the lower grades are inevitable. It is common to design mathematics examination papers by deciding in advance what proportions of questions should be accessible to different levels of candidate; the variation in grade boundaries is partly because this is not an exact science.

[30] Evaluating the adjacent levels model for differentiated assessment (2014), Tom Bramley
[31] Pearson, https://qualifications.pearson.com/en/support/support-topics/results-certification/grade-boundaries.html?Qualification-Family=GCSE

For teaching from September 2006, there were two tiers in GCSE Mathematics: Foundation and Higher. There had previously been three tiers: Foundation, Intermediate and Higher. The change to two tiers was made, at least partly, to ensure that all GCSE Mathematics candidates had the opportunity of gaining grade C – the threshold for a level 2 pass.

Ofqual summarised the differences in target grades between the three-tier and two-tier GCSEs in the 2004 and 2008 columns of the following table.[32] The allocated proportion of marks intended to be accessible at each grade for current GCSEs in Mathematics have been added in the third column.[33]

2004		2008 (teaching from 2006)		2017 (current) (teaching from 2015)	
Tier	**Grade allocation**	**Tier**	**Grade allocation**	**Tier**	**Grade allocation**
Foundation	G: 33% F: 22% E: 22% D: 22%	Foundation	G and F: 50% E: 20–25% D and C: 25–30%	Foundation	Grade 1 to lower part of grade 3: 50% Upper part of grade 3 to grade 5: 50%
Intermediate	E: 25% D: 25% C: 25% B: 25%				
Higher	C: 25% B: 25% A: 25% A*: 25%	Higher	D and C: 50% B: 20–25% A and A*: 25–30%	Higher	Grade 4 to grade 6: 50% Grade 7 to grade 9: 50%

Ofqual's comparison of standards between 2004 and 2008 identified the following consequences of the changes between these two grade allocations.[34]

> *Higher tier papers were less demanding because of the need to target 50 per cent of the questions at the lowest two grades in each tier. The positive outcome from this change, however, has been increased access for students at the lower end of each tier. It allows grade boundaries to be set especially for the C grade in the higher tier,*

[32] Ofqual, Review of standards in GCSE Mathematics 2004 and 2008, April 2012
[33] Ofqual, GCSE Subject Level Conditions and Requirements for Mathematics, May 2014
[34] Ofqual, Review of standards in GCSE Mathematics 2004 and 2008, April 2012

which allowed students to show what they knew, understood and could do. This was not always the case in 2004.

The change in demand of examinations for the first two-tier examinations resulted in higher grade boundaries at Higher Tier. Ofqual noted that this changed what A* students were expected to be able to do.[35]

The nature of what was being tested at A was being changed with extremely high levels of accuracy on easier questions being required as well as the ability to tackle the harder questions. What was lost was the test of thinking skills required by more complex unstructured questions.*

This change in demand led to concern from teachers about the transition between GCSE and A level. MEI surveyed teachers of A level Mathematics at the MEI conference in June 2009. Most of the teachers surveyed thought that the two-tier GCSE had made transition to A level Mathematics harder.[36]

It is clear from the responses that many teachers believe that the transition from GCSE to AS Level has been made more difficult for all students. Since this transition was already seen as difficult for many mathematics students, this could have serious implications.

Two factors would seem to be involved: the weakness of students' algebra and the structure of the examination papers.

There is nothing new in complaints about students' weakness in algebra on entering sixth form. The question is whether the problem has become significantly worse. There were many comments about the structure of the examination papers, each of which covers a large range of grades. It is clear from the responses that many people regard this as the root of the problem. There were many comments to the effect that the examinations, and the grades awarded, gave the students a false impression of their preparedness for AS and A Level.

With the (9 to 1) GCSE, first examined in 2017, there is a greater level of challenge at the highest grades than in 2008 but also lower grade boundaries at grade 4, compared to the boundaries for grade C. Higher tier GCSE Mathematics now covers 6 grades (with the additional possibility of grade 3 for learners who narrowly miss achieving grade 4); this raises the question of whether it is possible to set exam papers that appropriately challenge students throughout this grade range. It is usual for exam papers to start with the easiest questions and get gradually harder, but the current design means that they get too hard too fast for many students.

4.4 Technology in assessment

Developments in digital technology have changed the way in which mathematics is done – from everyday applications of numeracy to mathematical modelling and research. This is changing how and what mathematics is learned at school. The use of technology in assessment is lagging behind; consideration needs to be given to the role of technology in

[35] Ofqual, Review of standards in GCSE Mathematics 2004 and 2008, April 2012
[36] MEI, The effects of 2-tier GCSE Mathematics on transition to AS and A Level, July 2009

the assessment of mathematics. This is part of the drive for greater curriculum coherence in which the assessment assesses what is taught in a valid way.

This section is not about computer-based delivery of examinations, where students use the computer merely as the tool for answering their questions (as opposed to writing their answers with pen and paper). The focus here is the use of the technology in doing mathematics in examinations.

There are four possible routes to consider:

a. Skills and understanding gained from use of technology in lessons help students understand the mathematics even if technology is not used in examinations.
b. Some items in examinations directly advantage students who have used technology in learning (this has been attempted with the large data sets in AS and A level mathematics).
c. Incorporating more direct assessment involving technology use into a paper-based examination system or including suitable centre-based assessment as part of the final grade.
d. Computer-based examinations which include use of appropriate software applications.

Using technology in the assessment of mathematics allows direct assessment of the skills of using technology to do mathematics, including the use of appropriate software to explore real data. Computers have been used in MEI A level examinations since 2000 but only for papers with small numbers of candidates. This has allowed students to use software in examinations; however, the examination papers and answers are on paper rather than on screen. This is an example of c, above, as was MEI's Numerical Methods coursework.

The 2021 PISA assessments will be computer-based. The framework for assessment[37] outlines some of the affordances of using computer-based assessment of mathematical literacy. Adaptive assessment of mathematics at GCSE level could allow students to answer questions of an appropriate level of difficulty.

> *The main mode of delivery for PISA 2021 will be the computer-based assessment of mathematics (CBAM). The transition has been anticipated with both the 2015 and 2018 studies moving to computer-based delivery. In order to maintain trends across the studies, both the 2015 and 2018 assessments were computer neutral despite using a computer-based delivery mode. The transition to a full CBAM in 2021 provides a range of opportunities to develop the assessment of mathematical literacy to be better aligned with the evolving nature of mathematics in the modern world, while ensuring backward trends to previous cycles. These opportunities include new item formats (e.g. drag and drop); presenting students with real-world data (such as large, sortable datasets); creating mathematical models or simulations that students can explore by changing the variable values; curve fitting and using the best fit curve to make predictions. In addition to a wider range of question types and mathematical opportunities that the CBAM provides, it also allows for adaptive assessment.*

In late 2020, Ofqual published a review of barriers to greater adoption of online and on-screen assessment[38]. For some subjects, the use of technology in assessment may be

---

[37] Pisa 2021 Mathematics Framework (Draft), November 2018, OECD
[38] Online and on-screen assessment in high stakes, sessional qualifications. A review of the barriers to greater adoption and how these might be overcome. Ofqual, 2020

largely concerned with efficient delivery of assessments; for mathematics the use of technology would also allow more authentic assessments; practical difficulties regarding access to appropriate technology are similar in both cases. Ofqual's review came to the following conclusion.

> *It is clear, from the examples of New Zealand, Finland and Israel, that online and/or on-screen assessment can be successfully implemented in high stakes, sessional exams taken at volume in schools and colleges. This review aimed to establish a current view of the barriers to greater adoption of online and on-screen assessment in this context in England and to explore examples of how such barriers might be broken down and overcome.*
>
> *The review identified challenges to be overcome in schools and colleges including the lack of availability of sufficient devices, broadband and network capabilities in some cases, the variability in appropriate skills in teaching and support staff, challenges to overcome through implementation and the need to maintain fairness in delivering assessments digitally.*
>
> *The barriers identified were seen as real challenges to overcome by leading practitioners in the field, and representative groups of those integral to delivery of assessments in this context – teachers, headteachers, and others.*
>
> *Whilst the barriers identified are real, many are not unique to the circumstances in England. Each jurisdiction we looked at has taken a different path to implementation, making different choices as to how to manage the barriers in their specific circumstances to meet the needs of their students, qualifications users and broader education systems and to deliver the purpose and benefits the changes aimed for.*

4.5 Assessment of maths in context: pre-release material for mathematics examinations

Mathematical modelling and statistical problem solving always take place in a context which is important for the solution of the problem. Students may not be familiar with this context. When solving a problem in work or further study, they would become familiar with the context as part of working with the situation over an extended period of time. One way to give students familiarity with particular contexts in preparation for an examination is through pre-release materials. This is currently done for Core Maths examinations and also in A level Mathematics, via the Large Data Sets.

4.6 The advantages of timed written examinations

Having considered some of the features of examinations, we now consider some of their advantages. The next section considers some of their disadvantages, before non-exam assessment methods are considered.

Race[39], writing about assessment in the physical sciences in HE, lists the following advantages of timed written examinations.

[39] Race, P. (2009) Designing assessment to improve physical sciences learning. Higher Education Academy.

- ***Relatively economical.*** *Exams can be more cost-effective than many of the alternatives (though this depends on economies of scale when large numbers of students are examined, and also on how much time and money needs to be spent to ensure appropriate moderation of assessors' performance). However, any form of assessment can only be truly said to be cost-effective if it is actually effective in its contribution to students' learning.*
- ***Equality of opportunity.*** *Exams are demonstrably fair in that students have all the same tasks to do in the same way and within the same timescale. (However, not all things are equal in exams – ask any hay-fever sufferer, or candidate with menstrual problems).*
- ***We know whose work it is.*** *It is easier to be sure that the work being assessed was done by the candidate, and not by other people. For this reason, exams can be considered to be an 'antiplagiarism assessment' device, and although there are instances of attempting to cheat in exam rooms, good invigilation practice and well-planned design of the room (and the questions themselves) can eliminate most cheating.*
- ***Teaching staff are familiar with exams.*** *Familiarity does not always equate with validity, but the base of experience that teaching staff already have with traditional unseen exams means that at least some of the problems arising from them are well known, and sometimes well-addressed.*
- ***Exams cause students to get down to learning.*** *Even if the assessment method has problems, it certainly causes students to engage deliberately with the subject matter being covered by exams, and this can be worthwhile particularly for those 'harder' physical sciences areas where students may not otherwise spend the time and energy that is needed to make sense of the subject matter.*

Another advantage of timed written examinations is that the grading can be regarded as reasonably accurate, particularly for mathematics. Ofqual research finds that

> *The probability of receiving the 'definitive' qualification grade varies by qualification and subject, from 0.96 (a mathematics qualification) to 0.52 (an English language and literature qualification).*[40]

4.7 What can't timed written examinations do well?

Consideration of the issues raised so far in this paper might suggest some desirable features of an assessment which cannot easily be achieved in a timed written examination. These include:

- setting assessment tasks where the student can demonstrate positive achievement;
- the assessment of problem solving and modelling;
- the assessment of the use of technology;
- an assessment for which the purpose suggests a criterion referenced model is most appropriate;
- the assessment of collaborative working;
- the assessment of consistent performance and engagement rather than attainment on a particular day;
- the assessment of performance on sustained tasks;
- assessment using differentiation by outcome rather than by task.

[40] Ofqual 2018, Marking consistency metrics, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf

Exam-only systems also discard the teacher's knowledge of and evidence for the student's attainment.

The mental health of young people is a matter of national concern. Students' anxiety about examinations has been reported in the press. This is noted in an Ofqual review of research into anxiety about assessments.[41]

> *Exams and revision are known to be sources of concern for some students and this has likely been the case since their inception. Despite there being no measure systematically evaluating the degree of test anxiety over time, there is evidence that suggests that its prevalence has been relatively stable. However, anecdotal evidence from the education sector and perceptions in the media suggest that, more recently, children and young people are experiencing more stress and anxiety in relation to assessments. A recent and positive shift in attitudes towards discussing mental health may, in part, explain this.*

The review includes the following conclusion.

> *Overall, evidence suggests that the experience of test anxiety is largely a response to a disparity between beliefs of what can personally be achieved, the perceived difficulty of the assessment, expectations for performance and poor coping and emotion regulation. These beliefs can be set by the students, but can also be influenced by parents, peers and teachers. Ultimately, anxiety results from this disparity, which leaves the student feeling unprepared to manage the assessment, despite this not necessarily being the case in reality. Because of this the assessment, and preparation for it, is perceived as threatening, and test anxiety ensues.*

As well as the human cost on young people's welfare, anxiety could lead to underperformance in an examination, which might make the assessment less valid.

4.8 Non-exam assessment

The paper now considers various forms of non-exam assessment and some of their features.

Different types of non-exam assessment are possible in mathematics; these can be either a replacement for timed written examinations or in addition to them. The non-exam assessment might contribute to the final grade, as has been done with coursework in the past. Alternatively, it might be an additional endorsement, as in A level science for the practical work.

The types of non-exam assessment which have been used in mathematics include the following.

- Investigations
- Working with data
- Mathematical modelling
- Teacher formative assessment, with external checks; this is dealt with later in the paper

[41] Emma Howard, Ofqual, A review of the literature concerning anxiety for educational assessments, Feb 2020

4.9 THE ADVANTAGES AND DISADVANTAGES OF COURSEWORK FOR ASSESSING MATHEMATICS

The advantages of coursework address some of the disadvantages of timed written examinations. Coursework could:

- allow students to demonstrate positive achievement;
- better assess problem solving and modelling;
- assess the use of technology;
- assess using a criterion referenced model;
- assess collaborative working;
- assess consistent performance and engagement rather than attainment on a particular day;
- assess performance on sustained tasks;
- assess using differentiation by outcome rather than by task.

Coursework was a compulsory part of GCSE Mathematics at one time, and was part of some GCE Mathematics qualifications. QCA's review of GCE and GCSE coursework arrangements[42] states:

> *The review's findings confirm the value of coursework in many subjects. However, the report recommends that the assessment arrangements – including the role of coursework – for all qualifications should be kept under regular review. It also notes concerns raised by teachers about coursework in mathematics. The regulatory authorities will take full account of these concerns in their current development work on future mathematics specifications.*

In the main body of the report, it can be seen that concerns relate to authenticating coursework and to the data handling coursework component that was then part of GCSE Mathematics.

> *In contrast, 66 per cent of* ***mathematics*** *teachers indicated that coursework was sometimes problematic. GCSE mathematics teachers were concerned about authenticating candidates' work when formulae and answers were so readily available on the internet and because older siblings or parents could readily complete coursework tasks for candidates. The open-ended nature of the data handling exercise at GCSE left some candidates frustrated: there was no sense of completion since the exercise lent itself to continual development. The significant written element in this exercise was felt to disadvantage the candidates who were good at mathematics but poor at written English. The investigational project for mathematics coursework did, however, elicit some favourable comments from teachers and candidates alike.*

An oral component to coursework in the form of discussion or presentation can act as a check that the work is the student's own – this was done successfully in MEI A level mathematics coursework for many years.

Core Maths qualifications require a minimum of 80% of the assessment to be by external examination[43]; so it is possible to have up to 20% internal assessment. Only one specification has been accredited including internal assessment: the WJEC Eduqas Certificate in Mathematics for Work and Life. There were no candidates for the qualification, and it is no longer offered. There may be all kinds of reasons why teachers chose other

[42] A review of GCE and GCSE coursework arrangements, QCA, 2005
[43] Core maths qualifications: technical guidance, DfE, August 2018

specifications in preference to the Eduqas specification, but one possibility is that mathematics teachers continue to feel that coursework is problematic.

4.10 How to mark coursework

Another concern related to coursework is accuracy of teacher marking; constructing detailed mark schemes can increase agreement between markers but at the expense of narrowing the criteria and so failing to recognise good but unusual work. Wiliam[44] identifies a further issue – task-specific schemes can lead to teachers guiding students to the types of responses in the criteria but generic criteria can lead to the narrowing of tasks to those types which can be most easily assessed using the criteria.

> *To sum up, by delineating particular 'canonical' responses the task-specific schemes appear to lead teachers to direct students towards approaches that yield more easily 'assessable' responses. On the other hand, general schemes have tended to treat all tasks as equivalent, with scoring dependent upon the mathematical processes involved, which has placed a premium on selecting tasks that are likely to elicit the appropriate processes.*

An Ofqual review of the literature on the reliability of teacher assessment[45] reported that between 10% and 20% of centres had GCSE centre assessed marks changed at moderation but only 1% of centres had marks that were so inconsistent that a complete remark was required. Systems for implementing comparative judgement[46] could be used to produce a national rank order for coursework rather than asking each centre to assign marks to pieces of coursework.

4.11 Teacher assessment

Teacher assessed grades were used in 2020 and 2021 as a replacement for grades based on examinations. A different system was used in each year, both developed at pace in response to the crisis caused by the pandemic. Could a well designed teacher assessment system contribute to or even replace an exam-based system in the future?

Teachers assess their students as part of normal classroom practice. This formative assessment could be used as part of the overall summative assessment in some way. For example, it might be considered that teacher assessment could be used for assessing to a criterion referenced model and so might be a way of assessing a student’s understanding of ‘essential maths’.

Teacher formative assessment contributing to summative assessment also helps to meet the purpose of assessment in a way that exams do not.

- Part of the purpose of education is to prepare people for life and work. Success in life and work is not about cramming for a test, it's about consistent engagement and performance.

---

[44] Assessing authentic tasks: alternatives to mark-schemes, Dylan Wiliam, Nordic Studies in Mathematics Education Vol 2 , No l, 48-68

[45] A focus on teacher assessment reliability in GCSE and GCE, Sandra Johnson, Ofqual, 2011

[46] “Comparative judgement is an analytical process in which judges use their professional judgement to compare two scripts at a time, and to decide which of these scripts is ‘better’ in each case. Repeated comparisons result in a measurement scale showing the relative quality of the scripts.” (Pollitt, 2012).

- Some students do not perform well under exam pressure and so under-perform, meaning the exam result does not reflect the true level of their learning.

There is anecdotal evidence that students have been motivated to work hard to achieve good teacher grades in 2021.

The following would have to be set in place for teacher assessment to be used as part of a national assessment system.

- Agreement of criteria to be assessed in this way
- Examples of assessments - these could be questions to be used as part of normal teaching or short tests – or other acceptable forms of evidence
- Agreement of the conditions under which students should be assessed in this way, including what help they could, and could not, receive
- A system of record keeping and moderation which is not too burdensome and time-consuming
- Training of teachers and others involved

4.12 University entrance in Sweden

This final section considers briefly the approach to university entrance in Sweden to demonstrate that quite different approaches are possible from that in England.

Students in Sweden receive a teacher grade for each of their courses; they can gain entrance to university either as a result of their grade point average (GPA) across all their courses or by taking a test. Wikström[47] describes the situation as follows.

> *The Swedish situation is unique, by having two instruments for selection to higher education, very different in format and purpose, that can be used more or less interchangeably. While one instrument is a standardised test, the other is the result of teachers' decisions based on evidence of classroom assessment. The test is norm-referenced and the grades are aiming to be criterion-referenced. Both instruments have their problems and advantages. Grades are the result of evidence collected over a long period, while the test only represents a snapshot of a person's knowledge and skills on a certain occasion. On the other hand, the test can be taken an (almost) unlimited number of times. The grades include a varying degree of subjectivity, while the test is objectively administered and scored. Even though most students show a similar degree of success on both instruments, there are also categories of students who are advantaged by one and disadvantaged by the other.*

In particular, female students tend to be advantaged by GPA and male students by the standardised test.

Wikström reports that grade point averages have been rising steadily over time; she investigates a number of possible explanations for this:

> *Four hypotheses are presented as plausible explanations: better student achievements, student selection effects, strategic behaviour in course choices, and lowering of grading standards.*

She comes to the following conclusion.

[47] Criterion-Referenced Measurement for Educational Evaluation and Selection, Christina Wikström, Department of Educational Measurement, Umeå University, Thesis 2005

The results show that all grades are increasing in line with the GPA. When adding up the various results, the only plausible explanation is that grading standards have been lowered over time, which is interpreted here as grade inflation. The grade inflation is assumed to be an effect of the leniency of the grading system in combination with pressure for high grading, chiefly related to the upper secondary school grades' function as an instrument for selection to higher education.

She also compares grades in different schools and concludes that smaller schools and private schools tend to grade their students higher.

These schools attract high performing students, who are known to press for high grading and are also more likely to be more sensitive to competition than other schools. The conclusion is that grades are not strictly comparable among schools, which, among other things, negatively affects the grades' function as instruments for educational evaluation on school level, as well as student ranking.

□