Section 1: Scope, purpose and context of the consultation

If the reformed GCSEs are to meet the purposes stated, then key decisions need to be made in a timely and integrated way.

We are supportive of Ministers’ commitment to raising standards. For this reason, we are disappointed that there has been no early consultation on the performance standards of these new qualifications. Clarity and a shared understanding of the expected performance standards are important, from the outset, for the development of high-quality qualifications.

If GCSE results are to be used as an element of the school accountability system, then it is important to ensure that, in their design, they are aligned with the key accountability metrics\(^1\). If this is not the case, then there can be tensions between the ways in which standards for GCSEs are set and maintained and the accountability measures. For example, if the accountability system is intended to raise standards by encouraging schools to set and meet increasingly high targets, then a norm-referenced (ranking) approach to setting standards would not be the most appropriate one; indeed, as the reformed GCSEs will not be used primarily for the purpose of selection, there is little merit in a norm-referenced system.

If GCSEs are to be used to hold schools accountable for the performance of all their students, then it is very important that they should be appropriate for students across the ability range. An assessment is most reliable when it is targeted as closely to the level of ability of the student as is possible. For this reason, as we discuss below, it will be important, for some subjects, to retain tiering.

Section 2: Key design features – tiering

We support the principle that qualifications should only be tiered if appropriately demanding and manageable assessments cannot be developed without tiering. As we believe that this is only the case where differentiation is by task, then we support the principle that tiering should only be used where content exclusive to the higher tier can be identified.

The purpose of tiering is to support differentiation. An examination is most reliable when it is as closely targeted to the student’s level of ability as possible, so can test robustly at the threshold of what a student knows, understands and can do. Where examinations differentiate by task, tiering ensures that this is the case. The research evidence indicates that tiering within the examination itself is not necessarily a limiting factor on students’ ambitions. Decisions about what content it is appropriate to teach to particular students are often made at the beginning of a course of study, when setting decisions are made. Where tiering is used in an examination, research evidence indicates that, in a mature assessment system, teachers are able to make robust tier-entry decisions.2

Given that tiering is a response to a particular challenge in an assessment, it is AQA’s view that the default should be untiered assessments, with differentiation achieved by outcome. In this way, all students are given the same tasks, but respond in different ways and are graded appropriately. Unfortunately, not all subjects are amenable to differentiation by outcome. Some, such as mathematics, differentiate by task. Where this is the case, the approach to ensuring that the assessments are structured so that they differentiate most effectively and efficiently needs careful consideration. All students should have access to the “expected grade” – in the current GCSE, this is a grade C and an appropriate, and positive, test experience. The awarding bodies should be able to provide reassurance about the technical robustness of the approach and, in particular, that alternative routes to the “expected grade” are comparable.

There is a substantial literature on tiering3 and, in the current, mature system of assessment, the issues are well-understood. For that reason, AQA strongly urges that, for the delivery of live assessments for the reformed GCSEs, we build on current practice, rather than adopting new approaches, which carry substantial risk. Where there are new approaches which, prima facie, seem promising, given the stakes for students and schools, we urge that they are piloted and evaluated carefully.

With this in mind, AQA recommends the retention of tiering in mathematics. This is because mathematics most strongly differentiates by task, so it is essential that students are given

---

3 For example:
questions of the appropriate level of difficulty for them. If the questions are too easy, the test will be viewed as trivial; if they are too difficult, the students will not be able to access the test. In both cases, the test will fail to measure the student’s level of attainment accurately. The current two-tier (Foundation/Higher) model is well established, well understood and stable. Both tiers provide access to the threshold accountability grade and, with overlapping questions across question papers (questions which both Foundation and Higher tier students have been well-prepared to answer) we can use robust techniques, such as test equating, to ensure comparability between routes. It will, of course, be important to monitor test entries by schools, to ensure that accountability pressures do not predispose schools to enter students for the Foundation tier, in order to be (in their view) more confident of achieving a grade C. If there is evidence that students who are capable of achieving the higher grades are being entered for the Foundation tier, then that is an issue which will need to be addressed. The motivational aspects of the behaviour that could be driven in schools/colleges by the overlap in grade scales between tiers are central to its effectiveness. Whilst there are psychometric considerations which should be made when designing tiered assessments, these are related to the distribution of marks across assessments, rather than to the overlap of grades between the tiers. Whilst there is clearly a relationship between the two, it is likely to be such that the psychometric issues can be addressed through considered assessment design. Rather than designing the grading structure around the psychometric principles, it is, therefore, advisable to define the structure based on other factors and then consider how this can best be realised, from a psychometric perspective, during the assessment design phase.

The alternative approach, which is to use a core paper, which gives access to all grades and then additional papers to test the extremes of the grade distribution, might seem attractive. Our concerns with this model would be that it is untested and so could place substantial risk on the safe delivery of results in the early rounds of testing. There are further issues with this alternative, which would require substantial further investigation. For example, as candidates can take more than one paper, common items cannot be used. To ensure comparability, items need to be pre-tested. For a high stakes assessment, such as GCSE, there would be security concerns and pre-testing substantially increases costs.

Given the likely requirement, for the reformed GCSEs, to increase the qualitative demand of the questions (to offer students greater challenge in the assessments) and also possibly to raise the grading standard (so that the expected grade represents a higher level of difficulty than the current grade C), AQA very strongly urges that the approaches to setting and maintaining standards used are considered carefully and based on the most robust technical research evidence.

We recommend, strongly, that tiering be retained in both mathematics and science assessments and that the model used should be the two overlapping tiers model. We have indicated to Ofqual that it might be possible to develop a core plus extension model for science, but that recommendation was made as a response to a strong policy steer against the conventional tiering model. With the recognition of the need, in clearly argued circumstances, for tiering, we propose that the tried and tested model be used. We would, of course, be pleased to see research into alternatives, such as variants on the core plus extension model. We would, however, caution strongly against the use of alternative models in a high stakes, live examination without very robust piloting and evaluation.

We do not believe that the case for tiering in English reading is sufficiently strong to warrant a requirement for the reformed GCSEs to include it. We urge, however, that the performance of
lower ability students and, in particular, their ability to access stimulus materials, be carefully monitored and evaluated.

Section 3: Key design features: assessment arrangements

We are supportive of the default, that the reformed GCSEs should, wherever possible, be assessed by externally set and marked examinations, except where subject content cannot be validly assessed in this way.

So, for example, for GCSE English language, we are strongly supportive of the assessment of speaking and listening, as a key element of the subject. We consider it appropriate to support teachers in their assessments by providing exemplification and possibly moderation. We do not, however, believe it is appropriate for the outcome from the teacher assessment to be aggregated with that from the external examination into an overall grade. It is, in our view, more technically defensible to report the teacher-assessed grade separately, as an endorsement on the student’s certificate.

Where a final grade is based on externally set and marked examinations, we strongly support that there should be a minimum stipulated examination time and that three and a half hours seems an appropriate baseline for GCSE. It is important, we believe, for there to be flexibility and that awarding bodies are required to justify proposed examination times, in terms of validity/reliability, to Ofqual, in their Assessment Strategies.

We agree that reformed GCSEs can be linear, with examinations at the end of the course. May/June examinations, with a November examination session for resits in English and mathematics for students in year 12 and above, is appropriate.

As all examinations will be at the end of the course, we do not believe that, for GCSE, there need to be specific requirements about synoptic assessment. The intention behind the development of the notion of “synoptic assessment” was to mitigate the risk of a “learn and forget” culture developing in modular examinations for A level and ensure that, at the end of the course, students were required to integrate what they had learned and use that more fully developed understanding to answer questions relating to the whole of the syllabus. Where assessments are linear, there is an argument that, in the case of GCSE, they are inherently sufficiently synoptic to provide reassurance about a student's level of understanding of the whole subject.

Whilst we are strongly supportive of the policy intention behind the proposals for spelling, punctuation and grammar (SPaG), there is substantial evidence that approaches to assessing these key competences are very challenging. For GCSE English language, a 20% weighting might be appropriate. For other subjects, such as English literature, geography and history, the requirement for a 5% weighting for SPaG might have unintended consequences.

There are practical issues to overcome. In an item-based marking system, where a single examiner does not see a whole script, it is impossible for an overall SPaG mark to be allocated. The benefits, in terms of marking reliability, of an item-based approach are substantial4 and, for

---

that reason, it would not be prudent to abandon it, in order to allocate SPaG marks. Instead, SPaG marks are allocated to particular items, which is not optimal in terms of making an overall judgement about a student’s level of ability.

Historically, marks relating to spelling, punctuation and grammar – sometimes referred to as Quality of Written Communication (QWC) – have not always correlated well with those from the main construct being tested. So, for example, students who have demonstrated high levels of knowledge and skills in, say, geography, have not always performed so well at SPaG. In a recent report for Ofqual on the achieved weighting of assessment objectives, we found correlations between the QWC marks and the total marks in a GCSE Business Studies unit that ranged from 0.43 to 0.48. A more recent analysis of SPaG weightings in GCSE components found that correlations between total SPaG marks and total scores averaged 0.64.

There are two consequences of SPaG marks not correlating highly with the other marks on a test. First, it makes it virtually impossible to incorporate SPaG marks into levels of response mark schemes, where each level of the mark scheme identifies both subject-specific levels of performance and levels of performance in SPaG. Second, the SPaG marks fail to achieve their intended weighting, so have little bearing on a candidate’s final grade. In the aforementioned analysis, the SPaG marks achieved an average of approximately 75% of their intended weight, although there was considerable variation.

There is also the risk for borderline candidates. If, for example, there are two candidates on the A/B borderline and one gains a sufficient number of SPaG marks to be awarded a grade A and the other does not, a decision to select the candidate with the higher grade will not be based on his or her level of achievement with respect to the subject.

If the intention behind the requirement for a 5% weighting for SPaG is simply to signal its importance to students, rather than to measure performance in SPaG, then we can see some merit for the requirement.

Section 4: Key design features: reporting student performance

AQA strongly recommends that student performance in reformed GCSEs should be reported in terms of grades. Alternative approaches, such as reporting candidates’ total marks, standardised scores or national rank order provide alternative approaches, but would pose considerable challenges.

The use of candidates’ marks, rather than grades, to report performance risks suggesting false precision in the reporting of the outcomes of the assessments. Assessments have a finite reliability\(^5\), so there will necessarily be a difference between a candidate’s “true ability” and the marks they are able to accrue in any given assessment. If a decision were made to report marks, rather than grades, then additional statistical information reflecting the level of precision of those marks would have to be reported alongside those marks, in order to ensure that the marks were interpreted appropriately. Whilst, from a technical viewpoint, this approach would be robust, ensuring that users of the marks – such as teachers, admissions tutors and employers – could

---

\(^5\) Here, we mean reliability independent of marking-related reliability.
interpret them appropriately would be very challenging. There is a risk that inappropriate interpretations of the marks and/or of the additional statistical information about the limited precision of the marks could unnecessarily undermine public confidence in the assessments.

The reporting of standardised scores faces the same challenges and some additional ones. The purpose for standardising scores is to allow direct comparisons between marks to be made on different versions of an assessment; however, comparison of candidates’ standardised scores across years or across subjects is only valid where the distribution of candidates across the mark scale is closely controlled. Whilst substantial efforts are currently made to design assessments which result in stable mark distributions, such control is not achievable without the use of psychometrically-controlled tests, which require pretesting. This has substantial cost implications and, for a high stakes assessment such as GCSE, raises issues about security. Further, our research\(^6\) indicates that very careful consideration needs to be given to the use of psychometrics-based approaches for GCSEs, as the nature of the assessments does not always lend itself to these kinds of tests.

The reporting of candidate rank order is closely analogous to the reporting of standardised scores, although with less dependence on the precise shape of the mark distribution. It should be noted that, with multiple awarding bodies offering different versions of a qualification, making comparisons across specifications and across years adds an additional layer of uncertainty regarding the accuracy of any reported figure.

The continued use of classifications in the form of grades goes some way to mitigate the risk of over-interpretation of marks. Misclassification does, however, occur around the grade boundaries, as unavoidable unreliability in the marks achieved can lead to candidates with the same level of ability scoring just above or just below a grade boundary. These effects were investigated in the context of the current GCSE and AS units by Wheeldon and Stockford (2010), whose research demonstrated that the accuracy with which candidates were classified into grades inevitably varied between assessments. These variations were not necessarily a function of the quality of the assessments; rather, the ability profile of candidates relative to the position of the grade boundaries played a significant part in the accuracy with which candidates were, on the whole, classified. It is important to note that, whilst it is important for awarding organisations to monitor and fully understand the reliability of their assessments, it would not be appropriate for reliability, in this sense, to be a primary driver for the construction of those assessments, as this would pose a substantial threat to validity.

Increasing the number of grades in a grading structure necessarily reduces the accuracy with which the candidates, as a whole, can be classified; the more classifications there are, the greater the number of opportunities for candidates to be misclassified. When considering the introduction of additional grades, the tension between increased discrimination and decreased classification accuracy need to be borne in mind. There is some evidence, from the analyses carried out by Wheeldon and Stockford (2010) that the addition of a further grade on top of the eight grades currently available would not substantially reduce grade classification accuracy beyond its current level. Caution is required, though, if the intention is to increase discrimination in a particular area of the distribution – such as at the top end, between high achieving candidates – because this will lead to grade boundaries which are more closely spaced in this region of the ability scale. As Wheeldon and Stockford (2010) showed, the more closely spaced the grade boundaries are, the

---

lower the accuracy of the grade classifications. So a finer grading scale, intended to improve discrimination, can lead to higher levels of misclassification, due to the inevitable unreliability of any assessment.

Given their relative under-use, it would be reasonable for the number of grades available at the lower end of the ability scale to be reduced. It should not, however, be assumed that reducing the number of grades here would allow the number of grades at the higher end of the ability scale unproblematically to be increased. If a grade were removed from the bottom of the ability scale to allow an additional grade at the top of the ability scale, this would lead to a reduction in grade boundary separation, reducing classification accuracy. This risk could, in principle, be mitigated by making adjustments to the design of the assessments, but, to do this most effectively, would require a rigorous programme of pretesting which, as we noted above, is problematic for GCSEs.

In summary, then, decisions regarding the number of grade boundaries and their distribution across the ability range require a balance to be struck between the granularity with which it is useful to classify candidate performance against the level of classification accuracy deemed desirable. Given the complexity of the range of GCSE assessments, detailed, systematic modelling of the application of different grading structures and careful evaluation of the alternatives need to be carried out before final decisions are made.

Much of the discussion above has focused on the reliability of given assessments and it might, therefore, seem appropriate to argue that the issues raised could be addressed by the development of “more reliable” assessments. Whilst it is important for awarding bodies to monitor and fully understand the reliability of their assessments, developing assessments specifically so that they are optimally reliable can result in assessments where the content (knowledge) and skills assessed are narrow. Such narrowness is not appropriate for general qualifications and attempts to increase the reliability of GCSEs could place risk on the validity of those qualifications.

Changing the designation of classifications from lettered grades to a numerical scale, in isolation, is of no technical consequence. From a non-technical perspective, any change in the grading structure and/or associated recalibration of standards might mean that it is helpful to change the way grades are reported, to flag the change. It might also mitigate the risk of unhelpful comparisons being made between the outcomes from the current and reformed GCSEs. We understand that, in future, there will be no requirement for schools to monitor and/or report pupil performance at key stages 2 and 3 in terms of national curriculum levels, but there is, we believe, a risk that some users could become muddled by a numerical scale for key stages 2 and 3 which had the same numbers as a numerical scale for GCSEs.

We believe that it is very important for the grades for GCSEs in England to be easily distinguishable from those that will exist in the alternative qualifications offered in Wales and Northern Ireland.

Section 5: Full and short course GCSEs

We agree with the proposals for short course GCSEs. We believe that the standards of short course GCSEs should not be compromised by the requirement that they “fall out” of the awards for full course GCSEs. Further, we do not believe that large-entry short courses should be allowed to
drive the award of smaller-entry full course GCSEs. For these reasons, we strongly agree that short course GCSEs should be stand-alone qualifications.

Section 6: Regulating the reformed GCSEs

We strongly support the proposals for regulating the reformed GCSEs and the requirement for accreditation. We welcome the reassurance that the recognition process is likely to be less burdensome for existing GCSE awarding bodies, as they will be able to point to evidence from their existing activity. We would anticipate that Ofqual could operate a “light touch” recognition process for previously-recognised awarding organisations, only requesting additional information needed to demonstrate compliance with revised criteria.

In relation to our recent recognition to offer Project qualifications, we provided evidence of our suitability in terms of identity, constitution and governance, integrity, resources and finance and competence. We would not expect to have to resubmit all these documents again for future recognition submissions, as they have already been verified by your officers as being appropriate for your purposes.

In particular, we welcome the requirement that awarding bodies should develop assessment strategies for their reformed GCSEs. AQA already carries out robust systematic review of the effectiveness of our assessments and will continue to refine our approach. As with our current GCSEs and A levels, reformed GCSEs will be subject to such systematic evaluation.

Section 7: Subject-specific features of the reformed GCSEs

We have read the DfE subject content for English language, English literature, mathematics, sciences, geography and history and have made a full response to the proposals.

For English language, we strongly agree that the spoken language assessment cannot be assessed by an external written examination. The outcomes should not be aggregated with those from an external examination to form an overall grade, but reported separately. For those students who are granted an exemption, this should be reported on their certificate.

For English literature, we agree that the content can be assessed by externally assessed written examinations only.

For mathematics, we strongly agree that the content can be assessed by externally assessed written examinations only. We recommend strongly that the current two-tier structure is retained.

For science, we are of the view that the content can be assessed by externally assessed written examinations only. Practical skills should be assessed by the teacher, in accordance with awarding body requirements. The outcome should be reported as an endorsement on the certificate, as for speaking and listening for English language.

We strongly recommend that the practical science element should not contribute 10% to the student’s overall mark. This is because, for a practical assessment of this nature, discrimination between candidates will be limited and so the achieved weighting will be less than the 10% intended weighting. This will mean that the practical skills element of the assessment will contribute little to
the student’s final grade. We are strongly of the view that practical approaches to teaching science are essential to ensure student engagement and progression to A level and beyond. Nonetheless, assessment should not be the sole determinant of what is taught in schools.

For geography, we are of the view that the same arguments apply to fieldwork as to practical science. We see no reason why geography cannot be assessed appropriately through an external written examination.

For history, we currently assess through an external written examination and see no compelling reason for this to change.

Section 8: Equality impact assessment

We are supportive of the steps which Ofqual has taken to ensure a robust equality impact assessment and urge them to continue to work closely with experts in that field.