Awarding grades for the June 2020 examination series:

Qualifications Wales-regulated A-levels

Methods report

August 2020

Contents

Introduction

Background

Welsh Government announced the cancellation of summer 2020 examinations on 18th March, due to the ongoing COVID-19 public health crisis[1]. Following this announcement, Welsh Government issued Qualifications Wales with a direction that the summer 2020 cohort of GCSE, AS, A-level and Welsh Baccalaureate Skills Challenge Certificate learners should be issued with results this summer. The approach adopted had to be fair and robust, based on centres' judgements of their learners' attainment in each subject, and standardised using a range of other evidence[2]. 'Standardisation', in this context, is a process involving the use of statistical models to calculate grades.

Following a public consultation, Qualifications Wales confirmed their expectation that the models will produce grade outcomes for this summer that are *broadly similar* to previous years, and that candidates across the cohort should be awarded a set of grades that are, overall, a fair reflection of what they would have received had they sat exams. All candidates, including private candidates where possible, will receive a calculated grade.

This document sets out recommendations for the approach to calculating grades for **A-level** candidates. All A-level qualifications were awarded in the summer 2019 examination series, meaning that there are no new qualifications being awarded for the first time in summer 2020. A-level qualifications regulated by Ofqual but designated by Qualifications Wales for use in Wales are not covered by this technical report.

Key assessment principles

In selecting approaches to calculating grades for learners, we have been guided by the key principles of assessment – validity, reliability, fairness, manageability, and comparability.

Qualifications Wales defines **validity** as follows:

"*the extent to which the assessment tests the things it is supposed to assess. The use(s) of the outcome(s) of an assessment is/are valid if supported by evidence and theory. The evaluation of validity involves the development of a clear argument to support the proposed interpretation of the outcomes and the intended uses of the assessment. The validity argument should be built on statements of the proposed interpretation and supporting evidence collected from all stages of the assessment process".*

---

[1] Welsh Government (2020). *Cabinet Statement: Written Statement: Written Statement on Summer Examinations 2020.* https://gov.wales/written-statement-written-statement-summer-examinations-2020
[2] Welsh Government (2020). *Letter to Qualifications Wales from Minster for Education,* 6th April, https://gov.wales/gcse-and-level-cancellations-letter

This report is a form of validation argument. Ordinarily, validation arguments make claims about each feature or stage of the assessment process, to justify the judgements that we wish to make about a candidate's ability in the subject domain of interest[3]. As the grade calculation process is intended to replace an assessment process, here the validation argument is a justification of decisions made at each stage of calculation to date. The aim is therefore to demonstrate that – as far as is possible – the grades issued to candidates this summer will be as fit-for-purpose as the grades issued to candidates in any other examination series.

For general qualifications, the main purpose is to give a measure of learners' attainment that supports their future progression into work or further education[4]. This purpose has not changed. Under the heading of validity, therefore, the grade calculation process can be evaluated using the following principles.

Firstly, the **accuracy** of the approach used to calculate grades must be maximised. As we cannot know what grade outcome each learner would have achieved in this series, we need to use historical data to inform the choice of a model. In testing, we are likely to prefer models which most often correctly predict the grade that learners achieved via examinations, in a normal examination series. The choice of approach may differ for different qualifications, depending on what information is deemed to be most effective in correctly predicting outcomes.

Linked to accuracy is the need to deliver **fairness** for learners. In this examination series, we must ensure that any **bias** in the process of calculating grades against candidates with common attributes is minimised. Common attributes may include those covered by statutory equality duties, such as age, sex, race and disability; but also other considerations, such as socio-economic status, size of centre, and the language medium of entry. Considerations of bias are limited by data availability and quality, including data from previous examination series that can be used for comparison.

So that the goal of facilitating appropriate progression for learners is secured, grades must be **comparable** in meaning with previous series. Comparability of outcomes has two purposes in Qualifications Wales' definition: to ensure that fair comparison can be made about the attainment of learners with grades from different examination series and qualifications; and that outcomes can be used as a measure of standards in a subject over time[5]. In a normal series, the latter would be ensured via an awarding process and the principle of comparable outcomes, so that candidates are not advantaged or disadvantaged by any variation in challenge in the examination series in which they complete their qualification[6]. In this series, outcomes are expected to be 'broadly similar to previous series'. How this is achieved in grade calculation depends to some extent on the approach taken. This summer, some may also choose to compare approaches to calculating grades across qualifications and/or nations, to evaluate the extent to which grades are seen as comparable.

---

[3] Paul Newton (2017). *An approach to understanding validation arguments*. Coventry: Ofqual.
[4] Welsh Government (2020). *Letter to Qualifications Wales from Minster for Education,* 6th April, https://gov.wales/gcse-and-level-cancellations-letter
[5] Qualifications Wales (2020). Standard Conditions of Recognition. https://www.qualificationswales.org/english/publications/standard-conditions-of-recognition/
[6] Qualifications Wales, *A Closer Look at the Comparable Outcomes Approach*. https://www.qualificationswales.org/media/4806/comparable-outcomes-approach.pdf

The grade calculation process must also be **reliable**. Reliability usually relates to consistency – that each stage of the process would result in the same outcome if repeated. In a normal series, issues such as grade classification accuracy, quality of marking, and grade boundary checks are important aspects of reliability. When calculating grades, other aspects come to the fore, which relate closely to quality assurance and quality control:

- *Modelling considerations:* independent checks of model functioning.
- *Candidate-level considerations:* ensuring that candidates can only receive grades that they could reasonably have received in a normal series; accounting for likely resit improvement, early and multiple entry and banked assessment marks/grades; ensuring that candidate grade profiles across all subjects are similar to those seen in previous series.
- *Centre-level considerations:* plausibility checks of centre rank order, based on prior assessment data; reviews of centre outcome stability relative to previous series; accounting for known partnership arrangements between centres.

End-to-end, the grade calculation process should also be as **manageable** as possible for candidates, centres and those responsible for producing and quality-assuring the grades. Any reliable grade calculation process will place burdens on each of these groups, just as an examination series does. An unmanageable or overly burdensome process puts the intended purpose of calculating grades at risk, however. Models which are overly complicated or require too many data inputs may also create risks: parsimonious models have an advantage both in risk reduction but also in being more easily understood by learners, centres and wider users of qualifications and grades.

At all times, the recommendation and selection of the approach for calculating grades requires a balancing of factors relating to these key principles, but always focused on achieving the aims set out in the Welsh Government's direction and the requirements set out by Qualifications Wales.

Sources of evidence for calculating grades

A range of evidence is required for calculating and quality-assuring grades. Types of evidence can be grouped into six categories[7]. Under each category, the sources of evidence available for Qualifications Wales-regulated A-level qualifications have been reviewed to establish their potential value.

Marks/grades already obtained by candidates for assessments already completed as part of the qualification and **historical data about qualification functioning**

Where performance on one unit (for which a mark has been banked) is predictive of performance in another unit, it may be appropriate to either calculate marks or grades for assessments that candidates were unable to complete in June 2020, or for the qualification as a whole. Alternatively, the evidence could be used to quality assure centre assessment grades and rank orders.

Qualifications Wales-regulated A-level qualifications are based on a unitised structure, with unit assessments available on in the summer examination series each year. A majority of candidates study for the qualification over two years, taking the AS units in the first year, and the A2 units in the second year. AS units are weighted at 40% of the A-level qualification. Achieved marks are converted to a uniform mark scale (UMS) for each unit via the awarding process in each examination series. An example of entry patterns is shown in *Table 1*, for A-level History, for June 2020.

Table 1: Entry patterns for A-level History, June 2020 examination series

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Total
n	1314	603	4	35	24	1	1	18	2000
%	65.7	30.1	0.2	1.8	1.2	0.1	0.1	0.9	100.0

- Group 1 - Candidates who sat all AS units in June 2019 and all A2 units in June 2020 (no resits), 'cashing-in' (i.e. completing) the qualification in June 2020
- Group 2 - Candidates who sat all AS units in Summer 2019 and all A2 units in June 2020 (and resits in Summer 2020), cashing-in the qualification in June 2020
- Group 3 - Candidates who sat all AS and A2 units in June 2020 (no historical entries), cashing-in the qualification in June 2020
- Group 4 - Candidates who sat all AS units prior to June 2019, nothing in June 2019 and all A2 units in June 2020, cashing-in the qualification in June 2020
- Group 5 - Candidates who are resitting only this year, cashing-in the qualification in June 2020
- Group 6 - Candidates who sat all AS unit in June 2019 and A2 units in June 2020 (no resits), cashing-in the qualification in June 2020
- Group 7 - Candidates who sat some AS units in June 2019 and are sitting the rest with A2 in June 2020, cashing-in the qualification in June 2020
- Group 8 – Other routes

[7] This framework is derived from Ofqual (2020). Exceptional arrangements for assessment and grading in 2020: consultation on the assessment and grading of vocational, technical and other general qualifications.

There is therefore a significant volume of quality-assured banked evidence from previous assessments that could be used to inform the calculation of final grades, if the relationship between AS and A-level performance is strong. For those candidates cashing in the AS qualification in 2018, and then progressing to the A-level in 2019, there was a very high correlation between the total AS level UMS achieved in 2018 and the overall A-level UMS achieved in 2019 (0.846). This correlation is boosted somewhat by the inclusion of AS level UMS in the overall A-level UMS.

At the subject level, correlations between total AS and total A2 UMS are somewhat lower, ranging from 0.386 to 0.723 for qualifications with more than 500 candidates. This can be understood with reference to qualification functioning statistics that are produced routinely by WJEC for A-levels as part of an annual review cycle. In some cases, *unit* performance is highly correlated, although the picture differs markedly between units and qualifications. To illustrate this, *Figure 1* shows the unit correlations for A-level Music and A-level Biology. Correlations between unit marks for Music are generally low, reflecting the fact that each unit is designed to measure a different skill at AS and A2. Conversely, for Biology, unit performance generally correlates well with all other units. The nature of the construct and the structure of the assessment therefore matters when predicting outcomes, at unit or qualification level.

Figure 1: Unit correlations for A-level Music and A-level Biology, summer 2019

Music

Correlation matrix						
	1160U4_U5	1160U6_U7	1660U80	2660U10	2660U20	2660U30
1160U4_U5	1.00	-0.22	0.21	0.68	0.03	0.20
1160U6_U7	-0.22	1.00	0.31	0.01	0.43	0.29
1660U80	0.21	0.31	1.00	0.23	0.42	0.76
2660U10	0.68	0.01	0.23	1.00	0.15	0.25
2660U20	0.03	0.43	0.42	0.15	1.00	0.37
2660U30	0.20	0.29	0.76	0.25	0.37	1.00
Overall	0.48	0.46	0.86	0.51	0.56	0.82

Biology

Correlation matrix					
	1400U30	1400U40	1400U50	2400U10	2400U20
1400U30	1.00	0.80	0.61	0.75	0.76
1400U40	0.80	1.00	0.63	0.74	0.76
1400U50	0.61	0.63	1.00	0.58	0.61
2400U10	0.75	0.74	0.58	1.00	0.72
2400U20	0.76	0.76	0.61	0.72	1.00
Overall	0.92	0.93	0.73	0.86	0.88

Nevertheless, banked assessment marks are a highly trusted source of evidence for those candidates taking the predominant approach to the A-level qualification: completing AS units in the first year of study, and A2 units in the second year of study.. Two complicating factors need to be accounted for when using this evidence, however. Firstly, many candidates resit units. The prevalence of resit entries makes direct estimation more challenging. Secondly, a small number of learners will complete an A-level in one year, meaning that the candidate has no banked assessment marks. If banked assessment marks are used to calculate grades, these candidates may need a different approach to calculate a grade, depending on the overall method selected.

Internal assessment grades for completed work which is not yet externally moderated or verified

Non-examination assessment is a key element of many A-level qualifications. Before centres in Wales were formally closed in March, candidates may have completed some or all of the assessment, and in some cases, centres may have begun to mark and internally standardise completed work. Some centre moderation visits had taken place. Many candidates had not completed their assessments, however, and WJEC was not able to visit all centres prior to their closing or to fully quality assure the work of all visiting moderators.

Consequently, in order not to disadvantage candidates, and because the marks are not quality-assured, WJEC confirmed that it would not use provisional marks completed prior to the closure of centres in awarding grades[8].

Similarly, some oral and practical examiner visits also took place prior to schools closing; however, not all visits had taken place and therefore not all centres have been assessed .In order not to disadvantage any candidates, WJEC did not progress with the quality assurance processes for centres whose candidates had been marked. As these marks have not been quality assured, WJEC did not use these in awarding grades either.

Centre assessment grades (CAG) and rank orders

The Welsh Government direction requires that centres' judgements on candidates' attainment are used within the grade calculation process. As WJEC has not collected grade estimates for some years, there is no recent data to analyse to assess the accuracy of centre predictions. Including this data in the grade calculation process is important to validity as, in the absence of assessment performance evidence, teachers are uniquely placed to consider the attainment of their learners. There is evidence that teachers are better able to judge the attainment of their learners *relative* to each other than using an absolute estimate of the grades they will ultimately achieve (an *absolute* judgement), as teacher grade estimates tend to be positively biased[9]. For A-levels in Wales, the fact that most candidates will have completed AS units in a previous series means that most centres will have a strong initial basis for their judgements.

Centres' judgements have therefore been gathered in two forms.

- A *centre assessment grade*, based on what teachers would expect each candidate to achieve for each qualification, representing a fair, reasonable, and carefully considered judgement of the most likely grade that might be achieved in normal circumstances.
- A *rank order position* for each candidate within each grade. Centres were permitted to 'tie' candidates on a single rank, based on a scale dependent on the size of the centre's total entry. This is shown in *Table 2.* The scale was developed using a combination of statistical analysis and stakeholder feedback, based on the following principles:

---

[8] WJEC (2020). Coronavirus FAQs - Non-Examination Assessment (NEA).
[9] Tim Gill, Methods used by teachers to predict final A-level grades for their students. *Research Matters: a Cambridge Assessment publication*, 28 (Autumn 2019), pp.33-42

- we should not expect centres to be able to rank their learners to a greater degree of differentiation than the examination system would normally provide. In a normal examination series, centres will often see pairs or small groups of candidates achieve the same mark and grade;
- it will be necessary to limit the number of ties per candidate, so that the grade calculation process can function effectively – in particular to ensure that outcomes are broadly similar to previous series;
- based on analysis of data from previous examination series, there is a relationship between the size of the candidature and the number of other learners a candidate may share a result with within a centre. It may therefore be appropriate to allow more ties per candidate when a centre's cohort is larger.
- there was no need to extend the number of ties permitted beyond 15, even for the largest centres.

Table 2: Rules for allowing tied rank order positions

Maximum group rank size: candidates can be tied into groups of…	Number of candidates in cohort (all ages)	Maximum proportion at bottom of grade rank
No ties permitted	0-49	n/a
2	50-99	4.00%
3	100-149	3.00%
4	150-199	2.67%
5	200-249	2.50%
6	250-299	2.40%
7	300-349	2.33%
8	350-399	2.29%
9	400-449	2.25%
10	450+	2.22%

All candidates were ranked on a single rank order, regardless of age, for each subject. For A-levels, this information was gathered from centres at qualification level, between 1st June and 12th June 2020. Centres were given general guidance on how to collate and submit the required information[10], as well as subject-specific guidance on how to produce the grades and rank orders. This professional judgement was derived from evidence held within the centre (learner work or evidence of learner work), which had been reviewed by subject teachers/tutors/assessors and relevant heads of department. Private candidates were included in centres' submissions where centres believed that they had seen sufficient evidence to make a reliable judgement. In some cases, centres were unable to do this for particular private candidates and therefore centre assessment grades and rank orders were not collected for these candidates.

[10] Qualifications Wales (2020). Summer 2020 grades for GCSEs, AS and A-levels, and Skills Challenge Certificate (SCC) Information for Centres on the submission of Centre Assessment Grades https://qualificationswales.org/media/5973/information-for-centres-on-the-submission-of-centre-assessment-grades-version-2-18-may-2020.pdf

Heads of centres were required to confirm that their grades and rank orders had been checked for accuracy and represented a fair, objective and professional judgement of the grades that their learners would have been most likely to achieve in a normal series.

Analysis of centre assessment grades

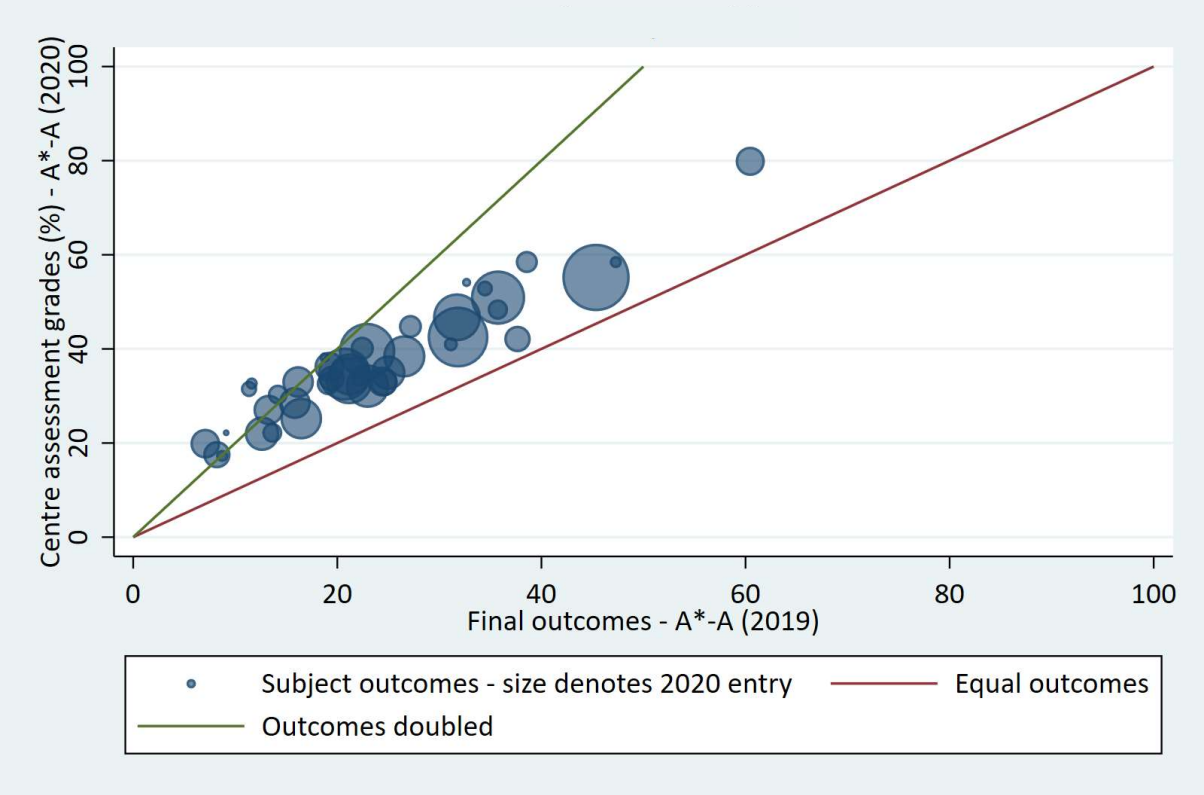
An analysis of the centre assessment grades provided by centres provides some insights into the quality of the data. As *Table 3* shows, overall, centre assessment grades (for Wales 18-year-olds) are much higher than summer 2019 outcomes were at the end of the period available for reviews of marking and moderation.

Table 3: Centre assessment grades in summer 2020, compared with final results in summer 2019 (Wales 18-year-olds only, cumulative percentages at each grade)

Subject (cumulative %)	A*	A	B	C	D	E	U	*n*
Summer 2019 (final)	**8.90**	**26.41**	52.11	76.78	91.75	97.88	**100.00**	*26315*
Summer 2020 (CAG)	**14.50**	**39.42**	68.03	90.18	98.12	99.92	**100.00**	*24169*

As *Figure 2* shows, the difference is reflected across all subjects at Grade A. A similar pattern is seen at all other grades.

Figure 2: Centre assessment grades in summer 2020, compared with final results in summer 2019, by subject (Wales 18-year-olds only) – proportion of candidates achieving a grade A or A* (cumulative percentage)



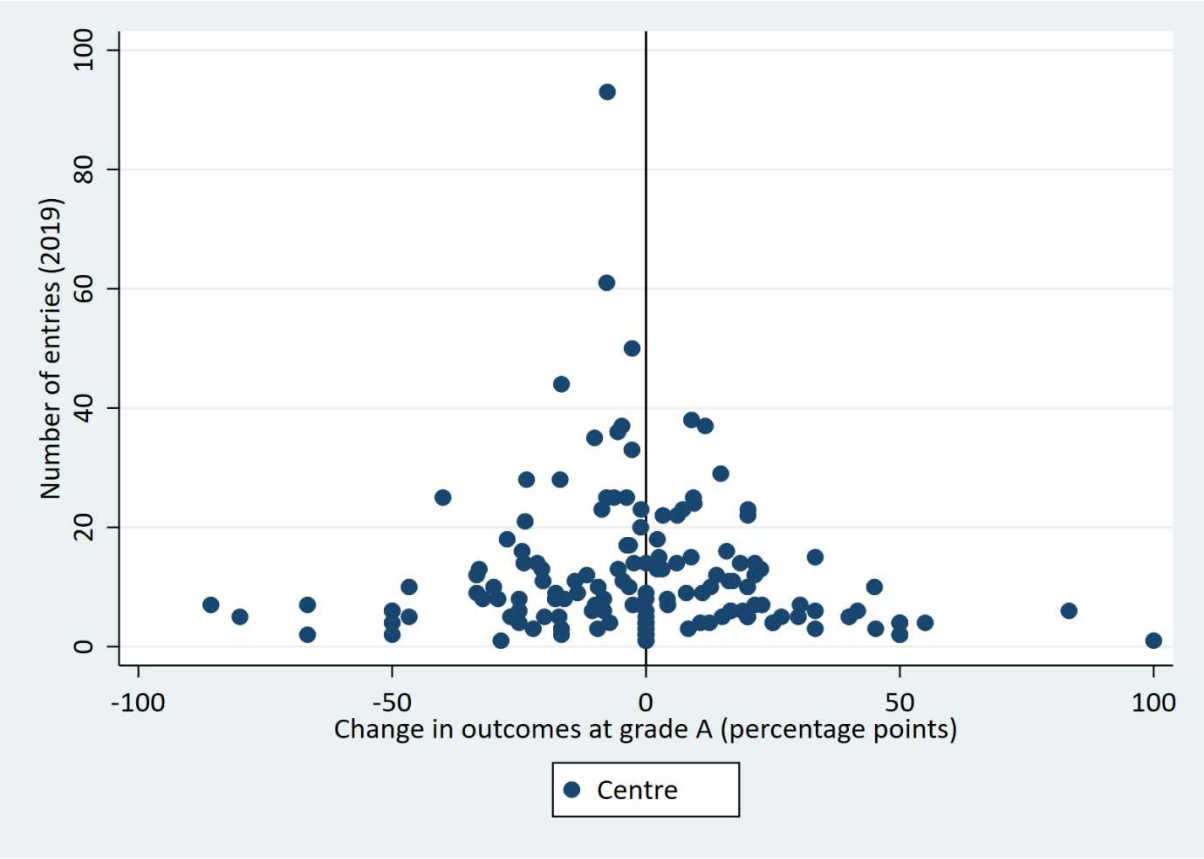
The grades and rank orders centres have provided on their learners are important, as – uniquely – they contain information on the attainment of learners throughout their current programme of study. The positive bias identified within the grades, which was expected, suggests that this information should be treated with care when calculating grades. Given the volume of information that was used by teachers in the process of determining rank orders, these can be considered to be more reliable.

Centre data and information

Centres' outcomes in previous series are a useful source of evidence in calculating and quality-assuring grades. Overall, we might expect that the pattern of variation in centre outcomes between series should be similar between 2018 and 2019, and between 2019 and 2020. At centre level, outcomes contain information about the performance of learners over time and the value-added provided by the centre. Centre outcomes may also differ by centre cohort size, with greater variation in outcomes amongst centres with small cohorts each year, or where cohorts differ markedly in size each year.

Figure 3 shows the centre-level variation at grade A for A-level Chemistry between 2018 and 2019. Smaller centre cohorts have outcomes which vary widely from series to series but, in general, centres with larger cohorts have more stable outcomes. Candidates' performance or attainment may differ, however, or the strength of a centre's cohort may differ from series to series – so some variation is to be expected.

Figure 3: Centre-level variation at grade A, for A-level Chemistry, between 2018 and 2019



Prior attainment data

Candidates' mean GCSE scores (based on candidates' grades in the summer of the academic year they turn 16) have been used as part of the awarding process for GCE qualifications for many years[11]. At candidate level, there is a moderately strong relationship (R^2=0.593) overall between Wales 18-year-olds' performance at A-level (expressed as UMS) and their mean GCSE decile; at qualification level, this varies between 0.30 and 0.85. Prior attainment data can give an indication of the strength of a cohort at centre or qualification level from series to series.

[11] *cf*. Benton, T. (2015). Can we do better than using 'mean GCSE grade' to predict future outcomes? An evaluation of Generalised Boosting Models. Oxford Review of Education, 41:5, pp.587-607.

Grade calculation options

In this section, the key options developed for calculating Wales A-level grades this summer are described. In order to tie the approaches with the analysis presented in this document, the outline of the procedure behind each of the methods is exemplified based on the modelling exercises undertaken using historical data, the results of which are presented in the next section of the report. Incorporating real examination outcomes served as the most reliable way of verifying the robustness of the methods, although it can never account for any issues of quality relating to centre assessment grades and rank order positions, or for issues of data availability in previous examination series.

After the estimated outcomes were calculated, different variants of ranking were initially explored for allocating grades, including scenarios involving tied ranks, unique ranks as well as ranks with an added random error[12]. The final testing was based on unique ranks generated using candidate UMS scores from the historical data set.

Direct centre performance (DCP) approach

The Direct Centre Performance approach is equivalent to the method developed by Ofqual to standardise grades for the general qualifications they regulate, which is based on a method proposed by Cambridge Assessment[13]. Candidate grades are based on a combination of calculated centre performance projections (based on historical centre performance data and information about the GCSE attainment of their candidates) and each centre's rank orders. The centre assessment grades themselves only influence the centre performance projections in terms of whether the estimated cumulative counts of candidates achieving each grade at each centre are rounded up or down, though an exception is made for small centres where – as described in the previous section – centre performance is more prone to variation over time.

For testing, grades from the 2016, 2017 and 2018 series was used to predict outcomes in 2019. The rank order was generated using the final UMS scores achieved by candidates completing the qualification in 2019. Two approaches to calculating grade probabilities were tested.

a. A log-odds version, whereby the weighted centre and national performance, and the centre predictions for the historical and current examination series were converted to log-odds values before the final calculation was made. The result was then converted back to a probability.
b. An alternative approach, whereby any centre performance projection that was calculated to be below zero or above one was adjusted back to zero and one respectively.

---

[12]The error was produced by generating random values using a model of a data distribution with mean=0 and standard deviation=5 based on the qualification maximum UMS. The random error was added to UMS scores that candidates received in 2019 and rounded to the nearest whole number.
[13] Benton, T. and T. Bramley (2020), *Estimating grades for candidates in GCSEs and A-levels in summer 2020 – Cambridge Assessment suggested approach*. Paper presented to Ofqual Standards and Technical Implementation Group (STIG), March 2020.

Once grades have been allocated to each centre, the grades were assigned to candidates based on the rank order.

Banked-unit centre performance (BUCP) approach

This method uses the sum of AS units' UMS scores to calculate outcomes for the cohort of candidates in a centre completing the qualification the following year[14]. Once the grades that the cohort in a centre should receive is calculated, these grades are then distributed to candidates according to the rank order of candidates provided by centres.

For each A2 unit, a z-score is calculated for each candidate using the sum of UMS AS units taken by candidates aged 17 (the predominant age group for AS), in order to give a standardised measure of performance. For any candidates re-sitting their AS units, an adjustment is added to their original UMS score based on the expected improvement to the grade, using the overall mean unit uplift calculated across all subject units. For testing, the uplift was based on data from 2017 and 2018.

The estimated A-level unit score are then combined with the AS unit scores (including the resit uplift where appropriate) to generate the total imputed score for every candidate. Cut-scores are set to produce overall outcomes for the cohort which match a target grade distribution (for testing, this is the 2019 result for each subject, for candidates included in this part of the model), so that grades can be allocated to the centre on the basis of the scores calculated for each centre's candidates. Grades are then allocated to candidates based on the centres' rank order.

Candidates not in the predominant age group and those without the required AS marks for inclusion in the main part of the model are 'slotted in' according to their rank order and centre assessment grade (CAG) and rank order position, so that a candidate receives the closest grade to their CAG which does not break the rank order.

Combined direct centre performance (DCP) with banked-unit (BCUP) approach

This method combines the elements of the DCP approach and the BUCP approach described above. Incorporating the z-scores produced by the BUCP approach in the calculation of the final outcomes using the DCP approach provides additional evidence of prior AS attainment that the original DCP method currently lacks. It was hypothesised that this could have a more direct relationship with the A-level grades than GCSE prior attainment and as such serve as a stronger predictor of the A-level outcomes.

For testing purposes, a z-score was calculated for candidates who completed their A-level in 2018, with a full allocation of AS units from 2017, and candidates were allocated into one of ten deciles based on their score. These were used to create a prior attainment-based prediction matrix of A-level outcome probabilities based on AS performance. This distribution was then used as a basis to calculate z-scores for the cohort entering AS in 2018 and completing their A-level in 2019. Predictions for each candidate were then created for each centre's A-levels in 2018 and 2019 using the prediction

matrix, and then aggregated to produce an allocation of grades for each centre. Grades were then allocated on the basis of centre rank orders. The essential difference between this approach and the DCP approach is that the predicted outcome used to create the centre grade allocations are based on AS performance in the subject, rather than overall performance at GCSE.

Mark-based regression (MBR) approach

In this approach, a multilevel model is created to predict a score for each candidate. These scores are converted into grades, and the grades are then distributed between candidates in a centre according to the centre assessment rank order positions provided. The boundaries can then be moved for these scores to get closer to prediction.

The initial model used a fixed categorical variable for prior attainment (deciles based on candidates' mean GCSE score) with a fixed linear effect for the AS UMS scored by candidates, and a random effect for centre. The model was 'trained' using the 2018 cohort (as well as 2017 for qualifications that were first awarded in that year), and then applied to the 2019 cohort candidates who were aged 18 and had a prior attainment score. Candidates not aged 18 or not matched to a mean GCSE score were 'slotted into' the model based on their centre assessment rank position, which for the purpose of testing was the order of candidates based on UMS achieved. The approach to "slotting in" for this model was that candidates were allocated a score equidistant on the UMS scale from the candidates above and below them.

Testing results and final model selection

The models were tested for a representative cross-section of A-level subjects, covering mathematics, science technology; Welsh and English Literature; creative arts; modern foreign languages; humanities and social sciences. The testing provided strong evidence for the BUCP method as being most accurate in predicting candidates' A-level grades. Using banked assessment outcomes from AS units as the primary basis for estimating A-level grades not only carries a high degree of validity, but also appears to more accurate than the use of GCSE-based prior attainment predictions. The balanced use of the statistical approach and the centre ranking that form the basis of the approach might allow centres and candidates higher confidence in the outcomes whilst allowing the grade standardisation process to remain as robust and bias-free as possible. The method showed no apparent differenced across gender and age when compared with the other approaches.

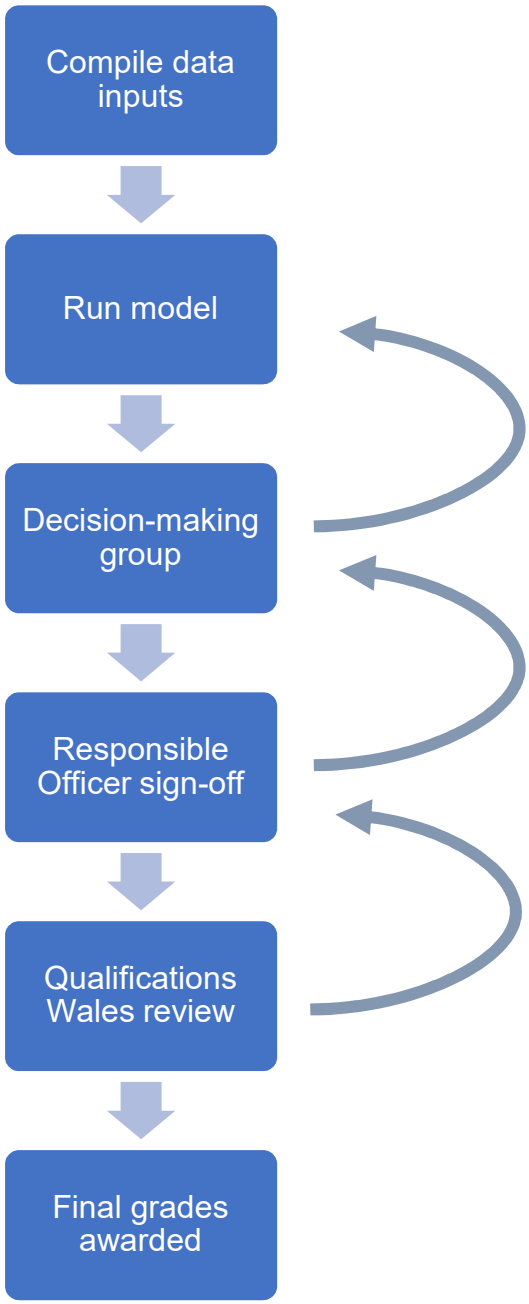
We recommended a common approach to calculating grades across the suite of qualifications for validity and manageability reasons, as the BUCP approach produced good accuracy outcomes (compared with actual 2019 grades awarded) across all subjects. This approach was approved by Qualifications Wales in July 2020. Centre-level variation was also measured for the selected model; patterns of variation were generally in line with those seen from examination series to series, with relatively few outliers.

The final approach

Process stages

Figure 4 sets out the approach taken to standardising grades for A-level qualifications. Each stage is explained in more detail below. Note that several of the stages are iterative. Depending on the decisions made at each stage, a qualification could be referred back to a previous stage. This could be in order to consider additional analysis, adjust the target outcome set within the model, to check data inputs, or to undertake additional quality assurance prior to the completion of the process.

Figure 4: Stages of the standardisation approach



Data inputs

Results data for Qualifications Wales-approved GCE qualifications was used as the first historical data input, in line with the requirements set down by Qualifications Wales in the Data Requirements for Summer 2020[15] document. This was combined with the centre assessment grades and rank order information provided by centres, as described previously in this report.

Prior attainment data for GCE qualifications is produced and quality assured annually by JCQ members for the purposes of setting and maintaining standards, on a three-country[16] basis, reflecting Qualifications Wales' Data Requirements. Candidates' mean GCSE scores are based on a common grade conversion which accounts for the fact that GCSE grade scales (9-1 for reformed Ofqual-regulated GCSES, A*-G for legacy GCSEs and reformed GCSEs in Wales and Northern Ireland) differed between qualifications in 2018, when the June 2020 18-year-old A-level cohort took most of their GCSEs. In the A-level model, this information is used only for the purposes of producing statistical predictions of the grade distribution of matched 18-year-old candidates in each subject.

Run model

Once all data was compiled and quality assured, the standardisation model was run.

Calculation stage

- In the first stage, for each 18-year-old candidate with all AS units banked from 2019, the sum of their AS units' UMS was calculated.
- For any candidates re-sitting their AS units, an adjustment is added to their original UMS score based on the expected improvement to the grade, using the overall mean unit uplift calculated across all subject units (the uplift was based on data from 2017 to 2018, and 2018 to 2019). Analysis regarding resit candidates indicated that candidates generally show a certain pattern of improvement based on the grade initially received when first taking the assessment. When this national figure (based on all subjects) is added to first-attempt grades, it reflects fairly closely the improvement in outcomes for the resit candidates. The uplift applied per unit (applied as a percentage of the original mark) was based on the following formula:

  New UMS value = Previous UMS achieved + ((uplift*maximum UMS achievable in the unit)/100)

  Uplift values, based on the original grade:

  A 2.437861
  B 7.079497
  C 8.807981
  D 10.37608
  E 13.31612
  U 18.04145

[15] Qualifications Wales (2020). *Wales Summer 2020 Data Requirements. GCE, GCSE, Welsh Baccalaureate Skills Challenge Certificate qualifications.* https://www.qualificationswales.org/english/publications/wales-summer-2020-data-requirements---gce-gcse-and-welsh-baccalaureate-skills-challenge-certificate-qualifications/
[16] England, Wales and Northern Ireland.

- The mean and standard deviation of UMS marks achieved by 18-year-old first-time entrants in each A level unit in 2017, 2018 and 2019 was calculated.
- The 2020 cohort z-scores were then combined with the A-level unit distribution parameters to produce an imputed score for each A2 unit. The AS z-scores were multiplied by an A2 unit's standard deviation (sd) and the outcome added to its mean (m): (z x sd) + m.
- An imputed score for each candidate was calculated by adding the (resit-adjusted) AS unit score to the estimated A2 unit scores.
- Note that although these are calculated using candidate-level UMS-based z-scores, the imputed scores are not assigned to candidates, but represent a set of scores for each centre so that the centre's rank order is protected.

Grade distribution and adjustment stage

'Cut scores' (the values at which scores are classified as representing one grade or another) are then set to assign all scores to a grade. It is possible to amend these values to bring overall outcomes closer to a predefined grade distribution. In the initial model run, the model parameters were set to award grades to a distribution based on principles agreed with Qualifications Wales.

- The cumulative proportion of the cohort achieving each grade should be higher than the cumulative proportion of candidates achieving that grade in 2019 after reviews of marking and moderation had been completed.
- The cumulative proportion of the 'matched 18-year-old' cohort (for whom mean GCSE prior attainment is available) achieving each grade should be higher than the statistical prediction defined by Qualifications Wales in the Data Requirements document for this series[17].

At subsequent stages of the standardisation process, target outcomes were adjusted to reflect the considerations of the decision-making groups, as well as the Responsible Officer, Standards Officer and Qualifications Wales.

Grade allocation stage

Once cut score values are set, grade allocations were produced for each centre, based on the imputed score distribution for their candidates. Grades were then distributed to the candidates included in the calculation stage of the model, based on the rank order provided by each centre.

Slotting-in stage

Candidates not included in the calculation stage of the model were 'slotted into' a grade according to their centre assessment grade (CAG) and rank order position, so that each of these candidates received the closest grade to their CAG which does not break the centre's rank order.

For example, for candidate X, if the candidate above X in the centre rank order received a grade B via the model, and the candidate below X receives an E, then if X's CAG is a B or better they will get a B;

---

[17] Qualifications Wales (2020). *Wales Summer 2020 Data Requirements. GCE, GCSE, Welsh Baccalaureate Skills Challenge Certificate qualifications.* https://www.qualificationswales.org/english/publications/wales-summer-2020-data-requirements---gce-gcse-and-welsh-baccalaureate-skills-challenge-certificate-qualifications/

if it is E or worse they will get an E and otherwise they will be awarded their CAG on the basis that it falls between the grades for candidates ranked either side of X.

Centres with no banked assessment evidence

In a small number of cases, no candidates in a centre's entry cohort has sufficient banked units to be included in the model. In these instances, the information available for standardising grades means that an alternative approach is required. The Direct Centre Performance approach used at AS was used to determine a standardised centre-level grade distribution, and grades were then allocated to candidates based on the centre rank order. A quality assurance check was applied in these instances to ensure that, overall, the qualification-level outcomes for the whole cohort would be aligned between the two methods.

An analysis of outcomes, setting out aggregated entries and proposed grade distributions, was then prepared for the decision-making groups to consider.

Decision-making group

A Calculation of Grades Subject-Specific Decision-Making Group was convened for each qualification being awarded in the summer 2020 examination series. Each group comprised:

- Responsible Officer (Director of Qualifications and Assessment Delivery) – Chair of meeting;
- Standards Officer (Assistant Director (Standards, Processing and Research));
- Assistant Director(s) from the Qualifications and Assessment Delivery directorate;
- Subject Officer(s) and/or domain leader(s) for the qualification(s) being discussed.

The purpose of the group was to review the approach taken to determining grades for each candidate entered for the qualification(s) being reviewed, and to decide if the approach is approved, or required further review or amendment. In making this decision, group members were asked to consider the following key principles.

- Validity and comparability – ensuring the grades issued in this series are fit-for-purpose, supporting appropriate progression for learners, that reflects their levels of attainment, and that are comparable in meaning to grades issues in previous series and by other awarding organisations where appropriate.
- Reliability – ensuring that the grade calculation process has sufficient quality controls and assurances in place.
- Fairness – that any biases in outcomes are minimised, so that learners with common attributes are not unreasonably adversely affected by the process of calculating grades.

Groups were either convened to meet remotely via Microsoft Teams, or virtually (in that no meeting was convened and feedback was gathered from group members for consideration by the Responsible Officer and Standards Officer once all group members had considered the proposed model and outcomes). Any personnel with a conflict of interest against any of the qualifications being discussed was asked to declare it before discussions commenced.

A Grading Partner from the Research & Standards team was responsible to preparing the statistical evidence for each qualification to inform the group, and then presented the key findings to the group. This evidence included:

- entries analysis: a breakdown of entries by age, gender, medium, country, centre type, private candidates, centre size, prior attainment, and other aspects where relevant.
- methods: an overview and justification of the statistical model used to calculate outcomes.
- grading outcomes: a comparison of cumulative grade distributions from the previous series, centre assessment grades (CAG), proposed final outcomes by the model. A breakdown by age, gender, centre type, and other groups was also provided where data was available.
- centre variation: including the change between cumulative CAG outcomes and cumulative proposed final grades at centre level for key grades.

Group members were then invited to provide their views on the approach and the rationale for the proposed outcomes, as well as any other concerns arising from the report.

Responsible Officer sign-off

The Responsible Officer, accounting for the statistical evidence presented and the feedback from Group members, decided either to:

- to accept the recommended approach and outcomes in full;
- to accept the recommended approach and/or outcomes, subject to a revision;
- to refer the qualification for further review by the Research & Standards department.

If qualifications were referred for further review, or a revision was requested, further evidence was presented to the Responsible Officer for further consideration prior to sign-off. If necessary, decision-making group members were invited to give additional feedback once the review was completed.

Qualifications Wales review

Grade outcomes were reported to Qualifications Wales as part of the Summer 2020 Data Requirements and discussed at regular bilateral Standards meetings between WJEC and Qualifications Wales, prior to final approval in August 2020. Once approval was received from Qualifications Wales, the final standardised grades were processed.

Final grades awarded

The grade outputs from the model were then assigned into grading systems. Several additional quality assurance stages were applied prior to issuing results.

- In a small number of cases, candidates had enough banked unit scores to cash-in the qualification and no unit entries. These candidates were awarded a grade based on the banked evidence, and their centres were asked not to include them in their centre assessment grade submission.

- In some other cases, where candidates were resitting units and cashing-in the qualification the standardisation model produced a grade for candidates which their existing banked evidence would have exceeded. These candidates were awarded a grade based on their banked evidence. This meant the rank order was broke in some instances, by agreement with Qualifications Wales, to uphold the principle of fairness. No candidate received a lower grade because of this adjustment.
- Private candidates with no centre assessment grade or rank order position were eligible to receive a calculated grade where they met criteria set down by Qualifications Wales[18]. As one of the requirements was to have completed all AS units in a previous series, a similar logic to have applied for the main grade calculation model was applied.

[18] Qualifications Wales (2020). Private candidate policy statement. https://www.qualificationswales.org/media/6184/private-candidate-policy-statement.pdf