The challenges of ensuring year-on-year comparability when moving from linear to unitised schemes at GCSE

Mike Forster Head of Research & Technical Standards, OCR

Introduction – the new GCSE suite

In September 2009, teaching began on a new suite of GCSE specifications. These were developed by the UK awarding bodies in order to meet new subject criteria specified by QCA. With the exception of English, mathematics, science and ICT (which were being developed to a different timescale), these new specifications covered all the available subjects. The outgoing specifications were mostly 'linear', whereby candidates took all the components of their assessment at the end of their course of study. The new GCSE specifications were all 'unitised' (also known as 'modular'). This meant that candidates could take units of assessment at different stages of their course, with the first units assessed in January 2010. As the nature of unitised and linear specifications is very different, it was imperative to ensure that appropriate unit grade boundaries were set, so that the first full course aggregation outcomes in June 2011 were appropriate.

OCR was in a unique position to be able to offer advice on this, in that it had been running a selection of unitised GCSEs for a number of years. Specifications in ICT, Business Studies, English and English Literature had been available with unitised assessment for up to nine years, and underwent little modification throughout the lifetime of the specifications. These were therefore the most appropriate specifications to use to illustrate the issues and difficulties facing those awarding (i.e. setting grade boundaries on) new units in the new specifications. Further background to the issue of modularisation can be found in D'Arcy (1997) and Vidal Rodeiro and Nádas (2010).

Comparability

There are many different definitions and contexts of comparability. In this article year-on-year comparability is the main concern. In an open letter to secondary schools and colleges dated 14th March 2011, Ofqual (2011) noted that the principle followed for the first awards of the new A levels in summer 2010, namely that there should be consistent standards at subject level between the old and the new specifications, would apply to the new suite of GCSEs in summer 2011. As such, Ofqual noted "we anticipate that the overall national results in summer 2011 for a particular subject will be reasonably close to the results in that subject in previous years". This particular definition of comparability is based on statistical comparability, as opposed to one based on expert judgement of the quality of candidates' work.

Thus the underlying imperative was to ensure that candidates of a given ability would achieve the same grade in the new specification as they would have done in the old (legacy) specification. In a time of relative stability, in comparability terms this would be what Newton *et al*. (2007) described as the "straightforward situation" of a parallel test, where the test is essentially assessing the same content in the same way, but with different questions. However, the restructuring of the new specifications and the change in the assessment model meant this was not that straightforward. Newton *et al*. also described a more complex situation – comparability of non-parallel versions of the test – in which the versions of the test are different in more than just the questions alone, for example with changes to both content and structure. It is this variant of year-on-year comparability that is central to this report.

Unitisation – the main issues

When linear specifications are awarded, the overall effect of component grade boundary decisions can be seen in the aggregation as a whole. If overall outcomes are not deemed appropriate, it is possible to revisit component boundaries until a satisfactory outcome is achieved. With unitised specifications, this is not possible in the same way. Aggregation outcomes for the new specifications were not available until June 2011. For any units taken before this date, decisions had to be taken whose impacts on overall outcomes would not be known until June 2011. Unit grade boundaries thus needed to be set with a view to satisfactory aggregation outcomes at a later date.

One advantage for candidates taking unitised specifications is the opportunity to resit units to improve their overall outcome. In the new unitised GCSEs, candidates were permitted one retake opportunity for each unit before requesting certification, and usually this would either maintain, or improve, overall grade outcomes, as the best unit outcome would be used towards the aggregation (subject to resit and terminal rules[1]).

One of the artefacts of the assessment of linear GCSE specifications in the UK is the use of two indicators for setting overall aggregation boundaries. The two indicators represent two possible boundary marks, and the chosen boundary is *normally* the lower of the two (as noted in Appendix 2 of the Ofqual code of practice (Ofqual, 2010)). Indicator 1 is the sum of the weighted boundary marks on each component, whilst Indicator 2 uses a weighted average percentage calculation to produce a boundary which yields an overall outcome more akin to the component outcomes. At the top judgemental grades (e.g. grade A on the Higher tier, and grade C on the Foundation tier) the Indicator 2 boundary is normally lower than the Indicator 1 boundary, with the opposite true for the lower grades (below the mean mark). This means, for example, that candidates could achieve a grade A overall (just!) by getting a top grade B on each component. In a unitised qualification, the overall specification boundaries are simply the sum of the unit uniform mark boundaries

1 The resit rules allow two attempts at each unit before certification (sometimes called 'cash-in'). Once a candidate has certificated, another two attempts are permitted before further certification, and so on. The terminal rule states that at least 40% of the assessment must be taken in the same session as certification. The marks from these units must contribute to the final mark, even if a better mark is available from an earlier attempt at the unit.

(see later for an explanation of the uniform mark scheme, UMS) – there is no Indicator 2. This meant that, unless allowance was made at unit level, candidates taking the new unitised specifications could have been disadvantaged at the top grades on each tier in comparison with candidates taking the old linear specifications[2].

The new suite of GCSEs included a 'terminal requirement'. This required candidates to take at least 40% of their assessment for a specification in the same session in which they aggregated (the process of aggregating all unit marks into one total, and receiving an overall grade – also called 'certification' or 'cash-in'). These units had to count towards the overall aggregation, and as such may have negated some of the benefit of resitting. It should be noted that no units were particularly designated as 'terminal units' – the terminal requirement could be met using any units (subject to meeting the 40% requirement and the resit rule).

A number of issues that could have affected unit outcomes needed to be considered, particularly in the first award of the unit. These could have made grade distributions appear odd, even if the outcome in grade terms was entirely appropriate. One such issue was candidate maturity. Candidates could take units throughout the course of study, and after as little as one term of teaching. As such, candidates who entered units early may not have had the maturity and knowledge of their linear counterparts. The performance required to achieve a given grade was the same regardless of the session of entry or maturity of the candidates, that is, the full GCSE standard was applied to all units. Assessors did not know the age of the candidates whose work they were marking, and no allowance was made for a lack of maturity. Therefore any lack of maturity or subject knowledge would have been evidenced by lower grade outcomes. This was especially the case for subjects such as modern foreign languages, where the nature of the cognitive learning process is more cumulative. It is also worth noting the difficulty in making judgemental decisions about the quality of work on units that assess (usually) smaller chunks of the specification, and add together in a different way from the legacy components.

An issue working in the opposite direction was that candidates could have gained an advantage (and hence improved their grade) as a result of the course assessment being broken down into 'bite-size' chunks, as they only needed to focus on one area of the specification at a time. Again, however, this benefit was constrained to some extent by the terminal rule, as candidates had to take at least 40% of the assessment at the end of their period of study. Ofqual's expectation was that there would be similar outcomes under the unitised scheme to those under the linear scheme. It was clear, therefore, that the pros and cons of the unitised scheme would to some extent cancel out, thus helping to ensure the structure of the assessment per se did not advantage or disadvantage the first cohort taking the new assessments.

Finally, centre entry strategies might also have produced misleading unit outcomes. Some centres might have entered only their most able candidates in the early sessions, whilst others might have entered all candidates to allow them to get a feel for what was expected. If the latter occurred in large enough numbers, the outcomes could have been very misleading indeed (even if they were entirely appropriate). Chairs of Examiners were therefore provided with age data about their cohort, which was used to support the awarding process.

Moving to a uniform mark scheme

This article has already identified a number of issues that could have had an impact when moving from a linear to a unitised specification. One such issue was the effect of introducing a uniform mark scheme, a necessity for a GCSE assessment that permitted units to be taken on different occasions.

As linear specifications assess candidates in one assessment window, the means by which candidates' scores are combined in order to produce an overall score is straightforward. When specifications are unitised, candidates can take units on different occasions, and it is therefore necessary to ensure parity between these units. This is achieved through a common mark scale – the uniform mark scheme. Raw scores are transposed onto a common mark scale such that equivalent scores (in terms of performance, not in terms of marks) from different sessions achieve the same number of uniform marks. Thus the candidate who scores the A boundary mark on a very difficult paper will get the same number of uniform marks as the candidate who gains the A boundary mark on a much easier version of the paper in another session. (See AQA, 2009; or Gray and Shaw, 2009 for further details). The mark transformations used in aggregating linear specifications are linear, according to the weighting of the components. In unitised specifications, the conversion rate for raw to uniform marks is not necessarily constant across the mark range[3]. This can result in compression or stretching of the raw-UMS conversion scale.

OCR replicated this effect by 'unitising' a number of existing linear specifications, to see the effect on grade outcomes. The outcomes varied from specification to specification, but there were some identifiable trends:

- On untiered specifications, most candidates tended to get the same grade following 'unitisation' as they had originally, but where grade changes did occur they tended to be downwards at the top grades, and upwards at the bottom grades.
- On the Foundation tier, most candidates tended to get the same grade as they had originally, but where grade changes did occur they tended to be downwards. On some specifications, this downward trend was restricted to the top grades, and on other specifications it was across all grades.
- On the Higher tier, as on the Foundation tier, most candidates tended to get the same grade as they had originally. Where there were grade changes, they tended to be downwards at the top grades, and upwards at the bottom grades.

These trends fitted with the expected impact of the removal of Indicator 2, namely that the proportion at the top grades would fall, but that at the lower grades the changes would be much smaller (or not there at all, if the boundary was set at Indicator 1). This supported the need to identify and act on the impact of the removal of Indicator 2 (see section below). However, there were also fluctuations at the bottom of the grade distribution, which suggested that subject-specific variations were also occurring.

2 Where the lower indicator is not chosen as the boundary mark (and assuming allowances are not made elsewhere), candidates at the lower grades on each tier who took a unitised specification would be advantaged over those who took a linear specification.

3 Between judgemental grades the conversion is linear (or quasi-linear if the intermediate gap is not equally divisible).

Unit versus aggregation outcomes

One of the major challenges facing awarders of the new GCSE specifications was setting new unit standards that would lead to acceptable specification outcomes when candidates aggregated for the first time. However, analysis of the existing unitised specifications showed little pattern between unit and aggregation outcomes. In some cases candidates gained the same grade on the units they had taken earliest as they gained when they ultimately aggregated. In other cases they did notably worse on the earliest units. Also, the introduction of new specifications (whether linear or unitised) invariably leads to a change in cohort, and most specifications also take time to find stability. As such, the outcomes on new units in new specifications may bear little resemblance to outcomes on the equivalent parts of old specifications. This was especially the case in some of the units taken in January 2010, whereby changes to the cohort in terms of centre type and age profile led to outcomes very unlike those seen in the legacy specifications.

To help ensure comparability year-on-year, OCR was able to account for the removal of Indicator 2 in the new unitised specifications by making unit-level adjustments to account for aggregate-level differences. Chairs of Examiners, who are responsible for the standards set in their awards, were presented with data which demonstrated the likely impact for each specification of the removal of Indicator 2. This was a basic numerical calculation showing the difference in the percentage in grade at the Indicator 1 boundary and the chosen boundary (usually Indicator 2). These impacts were then factored in to the awards at unit level to ensure overall outcomes were appropriate.

The issue of regression to the mean was also relevant. This is the situation in which component-level outcomes at the top and bottom grades are not maintained at aggregate level. Once aggregated, overall outcomes at the top grade tend to be lower than are found at component level, whilst overall outcomes at the bottom grade tend to be higher than are found on the components. The impact of this regression to the mean is determined by the correlation between each component (unit), and the number of components (units) in the assessment. If the number of components in the legacy linear specification was less than the number of units in the new unitised scheme, then there would have been greater regression to the mean in the unitised scheme, which would have affected overall outcomes.

Resitting pattern

The resit patterns for specifications that had previously been unitised (ICT, Business Studies, English, English Literature) were investigated. The resit rule permitted only one retake of any unit. The patterns for each unit, and each specification, varied somewhat. On the whole, the majority of candidates who resat a unit improved their mark, but this was not always the case. There appeared to be no overall pattern as to which session was the best to resit a unit in, with individual units showing different characteristics. The size of the mark changes varied too, although candidates who resat and improved their UMS mark tended to show larger mark changes than those who resat and got a lower UMS mark. This is not surprising since the resit candidates are a self-selecting sample, and few candidates would embark on their resit expecting to gain a lower UMS mark.

Year 10s and Year 11s

Unitisation offers the opportunity for candidates to sit a unit at any session in their course of study. The age profile of the cohort can have an effect on the outcomes for any unit. The majority of candidates who took the first units in the new suite of GCSEs in 2010 were from Year 10 (14 or 15 year olds), and as such their unit grade outcomes were below those seen in the components of the legacy linear specifications, which were mostly taken by Year 11 candidates (15 or 16 year olds). In comparison with the age profiles on the legacy specifications, this was much more variation in the age of candidates taking unitised specifications, and this could have affected the grade distributions for these units. To help with the setting of grade boundaries, Chairs of Examiners received the age profile of the cohort taking each unit.

Other statistical data

For new units in new specifications, the issues discussed presented a number of challenges to awarders. In summer 2011, these specifications certificated for the first time. Since achieving year-on-year comparability was paramount, it was possible to support the unit awards in this session with data about expected outcomes at specification level, based on measures of prior attainment, which Chairs of Examiners could use as one of the many indicators to help in setting appropriate standards. One of the main indicators for these specifications was a cohort-level prediction based on the average performance of the cohort at Key Stage 2 (KS2) – their average results in the national tests taken at age 11 in English, Maths and Science. This was achieved by establishing a matrix of mean KS2 performance against GCSE outcome for each subject, based on a historical pattern, and applying this pattern to the 2011 cohort for each subject to create a predicted outcome.

Summary

This article has noted the main issues that arose when the new unitised GCSE suite was examined for the first time in January 2010, and subsequently certificated in June 2011. The availability of resits, the loss of Indicator 2, the effect of maturity, the terminal requirement, and the introduction of a uniform mark scheme all had an impact on grade outcomes. The loss of Indicator 2 meant that, without correction at unit level, the proportion of candidates in the top grades (on both tiers) would have fallen. This was a fairly consistent finding. However, the other evidence was not so predictable. The analysis of unit outcomes against overall outcomes again showed a mixed pattern.

Data about resits on the existing unitised schemes also showed a mixed pattern. Most candidates improved when they resat a unit, but this was not always the case. Nor was there consistency in performance by session, with candidates in some specifications benefitting from a late resit (i.e. at the time of aggregation), whilst other candidates showed a similar improvement in each session. The make-up of the cohort by year group showed interesting outcomes, but again there was a lack of consistency. In most instances the Year 11 candidates out-performed the Year 10 candidates, but in some specifications the opposite tended to be the case.

The evidence from this article highlights the difficulties in accurately predicting what would happen when the new unitised GCSE

specifications were awarded for the first time. There were numerous factors influencing the outcomes, and whilst these were for the most part identifiable, there was no consistency in the patterns seen. What was true for one specification did not necessarily hold true for another. It was, therefore, crucial that Chairs of Examiners, and their awarding committees, used their judgement and experience, coupled with the statistical data available, to achieve outcomes that were comparable with previous years, and hence which did not give an advantage or disadvantage to the first cohort taking the new qualifications.

References

AQA (2009). Uniform marks in GCE, GCSE and Functional Skills exams and points in the Diploma. http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF Accessed 16/02/10.

D'Arcy, J. (Ed.) (1997). Comparability studies between modular and non-modular syllabuses in GCE Advanced level biology, English literature and mathematics in the 1996 summer examinations. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE. (Available on the CD rom which accompanies the QCA book.)

Gray, E., & Shaw, S. (2009). De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment Publication*, **7**, 32–37.

Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (2007). (Eds.). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Ofqual (2010). *GCSE, GCE, principal learning and project code of practice*. Coventry: Ofqual.

Ofqual (2011). *Open letter to secondary schools and colleges about the summer 2011 awards of new unitised GCSEs*. Coventry: Ofqual.

Vidal Rodeiro, C. & Nádas R (2010). Effects of modularisation. http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers