October 2012

Early entries and statistical predictions

Linear exams

In the past, students sat all of the assessments making up their GCSE at the end of a two-year course, in other words, *linearly*. Grade boundaries were set for each of the assessment components simultaneously using a combination of examiners' expert judgement and statistics (a process known as *weak criterion referencing*). This process allowed consistent qualification standards to be awarded year-on-year by taking account of the difficulty of the exams. The judgemental evidence has always been collected by scrutinising students' work and the statistical evidence is presented in the form of a prediction of the percentage of students expected to exceed each grade threshold.

The prediction takes account of how the students performed in earlier assessments and so reflects any changes in the ability of student cohorts between years. It is based on the concept of *value added* – a measure of the progress students make between different stages of education. This approach assumes the value added is constant no matter what group of students within the cohort we're looking at; the progress from their earlier educational assessments through to GCSE is assumed to be the broadly the same.

Modular exams

More recently the structure of GCSEs changed and most are now offered in a *modular* format. While the expert judgement used in standard setting is largely unchanged, modularity poses considerable challenges in the provision of statistical evidence, particularly when some modules are offered considerably earlier than others and a picture of the cohort's performance across the whole subject is still unknown. Module standards must be established with a view to maintaining standards across an entire subject in the future. To do this, we need to be able to answer a number of key questions.

Who's taking modules early and why?

The statistical predictions assume that the value added is constant between an earlier assessment and the GCSE for the students that are being awarded a *subject grade*. With the large and relatively stable entries for GCSE subjects this is a reasonable assumption. However, this assumption may not be valid when setting the standard for *single modules* taken early[1] on in the teaching of a course. The sample of candidates taking modules early is unlikely to be a random sample of those that will finally be awarded a subject grade.

In modular exams, the group of students entered for a given module is self-selected and the characteristics of this group are often very different to the cohort as a whole. This causes problems with providing robust statistical predictions. It is not uncommon in modular exams for students to be entered at a very early (possibly inappropriate) stage, and this is likely to affect their performance in the exam. There are many reasons why a student may be entered early, and most of them are things we simply cannot measure or take into account in our statistical

---

[1] Before the summer session of their second year of GCSE study.

AQA

predictions - school policy, parental pressure, candidate preference and so on. This problem is exacerbated if the module is part of a new specification and being taken by students for the first time. That means we have less to refer to in the form of previous students' performance in the exam in preceding years, and have to rely more heavily on expert examiner judgement and less on statistical evidence, because year-on-year evidence doesn't exist yet.

Because we don't know who they are and what motivates them, it is difficult to make appropriate modifications to our module predictions to account for early entry students.

How are the early entry modules contributing to the final subject standard?

A GCSE grade represents the sum of the assessment components or modules which contribute to the overall subject. In a *linear exam*, all these parts can be viewed concurrently which makes subject standard setting relatively easy. Grade boundaries for the *components* can be set whilst monitoring the impact on the overall *subject* outcome. In a *modular exam*, the parts can come from any sitting of the module and we need to anticipate how all the modules will fit together to achieve the required subject outcome. This again is particularly difficult if the module is part of a new specification and being taken by students for the first time, meaning we have less to refer to in the form of previous students' performance in the exam in preceding years.

The way in which modules combine depends on the extent to which candidates' performances on the separate modules are correlated. If the rank order of candidates is very similar on each module, the overall number of candidates achieving a grade will be very similar at subject level and at module level. Because of our remit to maintain standards year-on-year, we know roughly what the outcomes should look like across a subject as a whole; so if there is a strong correlation between the separate modules, it is easier to come up with a module prediction because it can be based on the expected subject outcome assuming, as discussed above, it is appropriate for the value added to be the same. Indeed, if the correlation between modules was perfect then the prediction for each module would be identical to the prediction for the overall subject (assuming the group of students taking the modules and the subject were the same).

In practice, exams are designed so that the modules are not perfectly correlated because they test different areas of the subject. A particular student may come out top in one module but be further down the ranking in a subsequent module. The correlation between the separate modules is therefore weaker, and the relationship between module and subject outcome is more complex and less predictable.

Because for early modules we can only estimate the extent to which candidates' outcomes on individual modules will correlate with one another, providing a statistical prediction for a module is challenging, particularly in view of the fact that decisions made when setting standards for early modules influence the future entry behaviour of students.

Achieving reliable and valid results

Exam boards work closely with the regulator, fellow awarding bodies, and large teams of researchers and statisticians to make sure our predictions and grade boundaries reflect the level of performance displayed by a student in a given exam. There are specific situations when the predictions we use become slightly less reliable, and early entry modules, particularly in their first sitting, are the prime example. However, we are not slaves to a single set of predictions; in these situations we gather a battery of bespoke evidence to inform the decisions that we make. Test equating using common questions across tiers and information about student age, student performance on concurrent early modules and centre performance, all help to ensure the standards set are reliable, robust, and students are rewarded appropriately for their performance in a given exam.