The case for scale scores - reporting outcomes in the reformed GCSE

Tom Bramley

25 April 2013

Research Division
Assessment Research and Development
Cambridge Assessment
1 Regent Street, Cambridge, CB2 1GG

Contents

1 Reporting scale scores

Cambridge Assessment is in favour of moving away from reporting results in the reformed GCSEs on a grade scale (e.g. A* to G) and instead reporting results on a numerical scale with a much larger number of points. This would have the following advantages:

– It would preserve as much information as possible about the rank order of students obtained from the examination.

– It would avoid the situation where two people can have scores some distance apart yet receive the same grade while another two people can have scores very close together but receive different grades because they are either side of a grade boundary. See Figure 1 below.

– It would mean that unavoidable measurement error would have a similar impact along the scale, rather than its effect being concentrated around the grade boundaries.

– A longer numerical scale should be more 'future proof' in that there would be no need to add extra grades (A*, A** etc.) if achievement rises and/or more discrimination is needed at the top end.

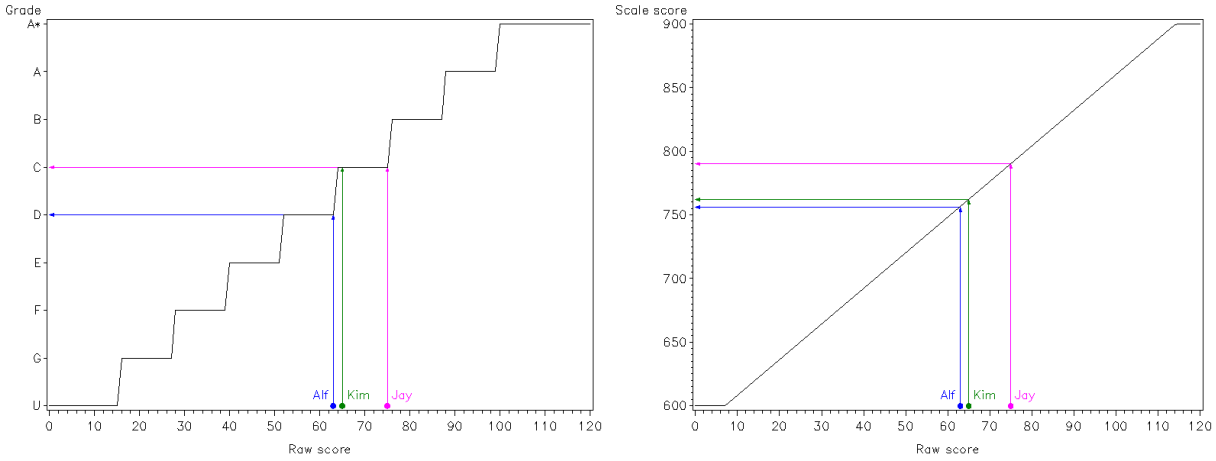

Figure 1: Mapping raw scores to grades (left) and to scale scores (right).

The graph on the left shows how raw scores might be mapped to grades (in the simplest possible case of an un-tiered linear exam). It can be seen that Alf and Kim, whose raw scores are close together, end up with different grades while Kim and Jay, whose raw scores are further apart, end up with the same grade. The graph on the right shows how the same raw scores might be mapped to numerical scale scores. The mapping preserves the information that the distance between Jay and Kim is greater than that between Kim and Alf.

The numerical reporting scale should be chosen such that there is no possibility of it being confused with a scale based on 'percentage of marks gained'. For example, scale scores could range from 600 to 900 points. There should be more available values on the reporting scale than on the raw score scale in order to

ensure no loss of information (i.e. to avoid many-to-one mapping of raw scores to scale scores[1]). The transformation from aggregate raw score to scale score should be as smooth as possible over most of the raw score range. This would avoid sudden changes in the value of one more raw mark.

For example, in the simplest case – an un-tiered linear syllabus – the raw-to-scale score transformation could be defined by identifying an upper and lower anchor point on the aggregate raw score scale which would map to fixed values on the reporting scale, and linearly interpolating or extrapolating all other raw scores (in a very similar way to the Uniform Mark Scale (UMS) currently used for modular A levels and GCSEs).

[1] It may be necessary to set a 'ceiling' such that scale scores above a certain point are reported as the ceiling value, and it may be necessary to set a 'floor' such that scale scores below a certain point are reported as the floor value.

2 Questions and issues

2.1 People understand what grades mean, but won't be able to interpret a scale score

It is not clear in the current system that grades can be (or are) interpreted in a criterion-referenced way – that is, in terms of what pupils with a given grade know and can do. In a sense, the grades are arbitrary categories imposed on an underlying continuum of achievement. Reporting scale scores would more faithfully capture this idea of a continuum. Meaning could be given to the scale scores by contextualisation – for example, in a criterion-referenced way by providing examples of questions that pupils with a given scale score are likely to be able to answer correctly or incorrectly; and in a cohort-referenced way by reporting the proportion of the cohort achieving at or above each scale score.

2.2 How would tiering be handled?

The issue of differentiated assessment raises problems whether grades or scale scores are reported. Tiered GCSE papers currently have overlapping grades available (C and D, plus a restricted E range available on the higher tier to prevent higher tier examinees who just fail to get a D being ungraded). Analogously, overlapping ranges of scale scores might be available to examinees taking different tiers of an assessment. However, the attempt to assure the comparability of scale scores in the overlapping range may expose (N.B. not create) the technical problems inherent in 'vertical scaling'. Therefore a move to reporting scale scores may also involve a reconsideration of what models of differentiated assessment will work best to allow the reformed GCSEs to be as fit for purpose as possible. Not all forms of tiering involve overlapping grade ranges.[2] It may be that a tiering model that provides only one route to a given range of scale scores would lead to more secure inferences about what examinees know and can do.

2.3 How would standards be maintained?

In principle the same problems would arise as in a grade-based system, and similar methods could be used to address them. Clearly there would be a role for the regulator in requiring the different exam boards to use the same reporting scales, and in ensuring that their procedures for transforming raw scores to scale scores (the analog of grading) could support claims of comparability across boards and over time.

Of course, specific technical challenges and opportunities relating to the longer and more fine-grained scales would arise, and these could be dealt with by the usual channels for inter-board collaboration and sharing of expertise.

2.4 How would the new numerical scale relate to the old grade scale?

This would depend on how much desire there is for the reformed GCSEs to make a clean break with the past. In principle the new numerical scale could be created such that certain scale score ranges mapped to old grades. But in the context of reforms to both syllabus content and assessment structure it might be

[2] See for example Baird, J., Fearnley, A., Fowles, D., Jones, B., Morfidi, E., & While, D. (2001). *Tiering in the GCSE.* London: Joint Council for General Qualifications.

argued that inferences about the capabilities of examinees from the old and new GCSE would not be very secure. It may be better to design the assessments such that they can provide a sensible mapping from raw score to scale score, and then simply to note the scale scores corresponding to 'comparable outcomes' with the previous system.

2.5 What might the consequences be of a shift from grades to scale scores?

Reporting scale scores might encourage the use of different accountability indices (e.g. means and standard deviations) to describe performance at aggregate level, instead of the focus on 'percent achieving above grade C'. This in turn might reduce some of the undesirable effects in schools of extra effort being concentrated on pupils around the grade C boundary.

Build-up of experience with a more fine-grained scale might allow different stakeholders to decide on different cut-off points on the scale to suit their own purposes.