EXTENDING THE NUMBER OF GRADES ON A QUALIFICATION: ISSUES TO CONSIDER

Lesley Meyer

1. BACKGROUND

The issue of the number of grades a qualification offers is currently a hot topic; University staff are complaining that they are unable to discriminate between the most able candidates when considering applicants to their courses and there are claims from some quarters that the International Baccalaureate (Diploma Programme for ages 16-19) provides better discrimination that the A level qualification because of its different grading system[1]. Reform to the A level qualification therefore seems likely and it may be that part of that reform will be to increase the number of grades the A level supports. This paper has been prepared to assist discussions about the implications of such a reform[2].

The number of grades used to report examination results varies considerably. In UK qualifications covered by the examination regulators' Code of Practice alone, GCSE examinations use nine points (A*, A, B, C, D, E, F, G, U); A levels use seven (A*, A, B, C, D, E, U); FSMQs[3] use six (A, B, C, D, E, U); and the Principal Learning and Project qualifications use seven at Level 3 (as per A level), five at Level 2 (A*, A, B, C, U) and four at Level 1 (A*, A, B, U). Accordingly, it is clear that there is no generally accepted rationale for deciding the number of grades which should be used to report examination results. Indeed, there are various issues to consider: some of these relate to the reliability of the underlying mark scale and usually imply adopting a relatively small number of grades; others stem from the loss of information incurred when a small number of relatively coarse categories is used, justifying the use of a larger number of grades. Overall, the number of grades used to report achievement on any given examination depends on the relative importance of many different factors, which are outlined below (in no particular order).

2. INCREASED LIKELIHOOD OF MISCLASSIFICATION

If the number of grades on an assessment is increased (without increasing the total mark on the paper) the implication is that the grade boundaries will get closer together, i.e., the width of each grade range will narrow. However, the closer a learner is to a grade boundary the greater the chance that he/she will be misclassified; in the middle of a grade boundary the chance of accurate classification increases (Stockford, 2011), Figure 1.

[1] Diploma Programme students follow six courses (i.e., six individual subjects) at higher level or standard level. The grades awarded for each course range from 1 (Very poor) to 7 (Excellent), or N if the result is no grade. Students can also be awarded up to three additional points for their combined results on two additional components: theory of knowledge (TOK) and the extended essay. The latter two components are graded from A (Excellent) to E (Elementary) or N (no grade). Overall therefore, the highest points total that a Diploma Programme student can be awarded is 45 points. The Diploma is awarded to students who gain at least 24 points, subject to certain minimum levels of performance across the whole Diploma and to satisfactory participation in creativity, action and service (CAS). The results indicate the grade a candidate has been awarded for each subject, including the TOK and extended essay components; they also indicate the completion of CAS and the total points for the diploma, if a diploma has been awarded.

[2] There are likely to be other assessment related factors to consider, for example there may also be changes in the demand and content of the A level material; however, this paper focuses solely on the implications of increasing the number of grades.

[3] Free Standing Mathematics Qualifications

So, if the grade boundaries are closer together because there are more grades, then the number of misclassifications is likely to increase. Willmott and Nutall (1975) summarised this picture succinctly:

> *Moreover, the more grades that are used, the more the number of candidates misclassified rises but the more the severity of each misclassification falls: the fewer grades that are used, the fewer the number of misclassifications but the greater the severity of each misclassification so made. It is a neat situation…*

Of course, it may be considered that the fact that the severity of the misclassifications is reduced by increasing the number of grades outweighs the disadvantage that fewer candidates will receive their actual true grade. Ultimately though, if the number of grades on a qualification is extended, a possible implication is that the number of raw marks on each paper could be increased to enable wider grade widths. In a modular system (or any examination system using the UMS), the UMS scale would also need to be extended.

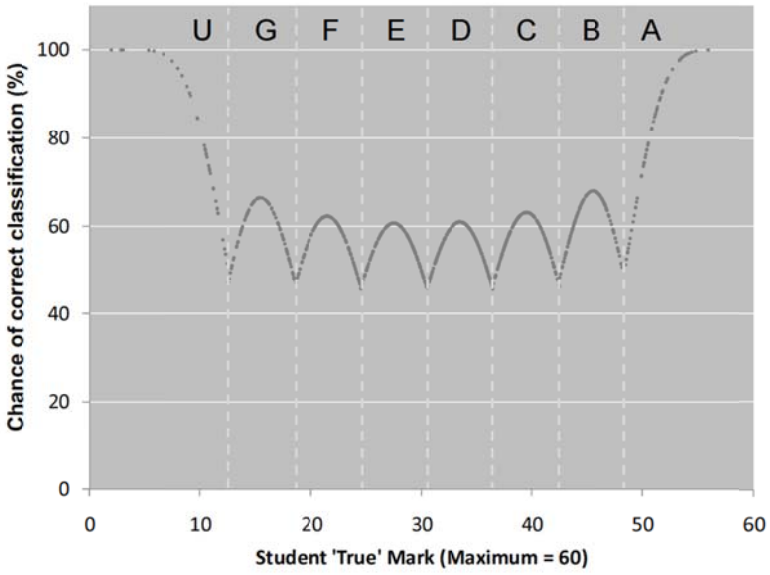

Figure 1 **Chances of candidates of different ability being correctly classified on an exam with multiple grade boundaries**

In theory, increasing the length of a test increases its reliability and the chance of correct classification (in Figure 1, the gradient of the humps between the grade boundaries would steepen). However **the burden of assessment is also then increased**; if a test becomes too long (or the aggregate of the total testing window becomes too burdensome), learner performance may be affected as learners may become tired or bored, and may therefore stop performing to their best ability. Thus increasing the burden of assessment should not be taken lightly and any theoretical gains should be set against these practical considerations, particularly as the extent to which test length affects the number of grades that can be supported is complex and dependent on a number of other variables. For example, doubling the length of the assessments does not imply that the number of grades the qualification can sustain will double and other factors, such as the number of components being aggregated and the correlation between those components, will affect the relationship.

3. SHOULD THE VERY TOP AND BOTTOM GRADES BE JUDGEMENTAL?

In all examination types, certain grade boundary marks are established according to a combination of the awarding committee's judgement and the statistical evidence; these are termed 'judgemental' grades. The remaining grades, the 'arithmetic' grades, are calculated by extrapolation and interpolation. On an arithmetic grade, the standard of work associated with the grade may change year on year, whereas on a judgemental grade the standard is held (approximately) constant, having adjusted for candidates' prior ability. The required standard of work at the judgemental grades is outlined in the subject criteria[4].

In general, awarders find it easier to make judgements at (or near) the ends of the range of attainment rather than towards the middle, which would support the two extreme boundaries being fixed judgementally (Cresswell, 1986). Further, if the root of the reason for any proposed extension to the number of grades available in a qualification is to increase the differentiation at the top end, heavy emphasis will be inevitably placed on (particularly) the highest grade. In that sense, there would be an argument to award the highest grade by a combination of judgement and statistics (i.e., for it not to be an arithmetic grade).

However, at the top end of the grade scale, there may be a paucity of work available for scrutiny which would hinder a judgemental approach, potentially decreasing the reliability of the examiners' judgements at that grade. This would be of particular concern if they were reliant on judgement alone (for example, in an examination with a very small entry for which statistical evidence was limited or not viable)[5]. Therefore customers' emphasis on the very top grade may be misplaced, as if the grade is calculated arithmetically the standard of work may change year on year (and also is not associated with any subject grade criteria), and if it is established judgementally the judgemental evidence to support the grading decisions may be limited. The greater the number of grades the more this intensifies, as the top end becomes more extreme.

This has particular resonance when considering A level examinations, in which currently grades A and E are judgemental and the newly introduced grade A* is calculated arithmetically. Whether grade A* should replace grade A as the upper judgemental grade is already being debated; if consideration was being given to extending the number of A level grades, the discussion about which top end grade (or grades) should be judgemental would intensify – particularly given the weight that will be placed on those top end grades.

4. SHOULD THE NUMBER OF JUDGEMENTAL GRADE BOUNDARIES BE INCREASED?

As the number of grades increases, allowing greater numbers of grades to be arithmetic and 'float' in terms of standards year on year, may become intuitively less acceptable. However, increasing the number of judgemental grades will lengthen awarding meetings and stretch resources. Also, grading judgements are difficult to make. It is therefore desirable for awarders to be asked to focus on only a limited number of grade boundaries and, more particularly, only those for which they are capable of conceptualising the standard. Moreover, the greater the number of grades the more important it is for

[4] Specifically within the grade descriptors (at GCSE) or performance descriptors (at A level), which are part of the subject criteria.

[5] Pinot de Moira (2008) showed that the nearer a prediction is to the extreme ends of the grading scale the tighter the percentile interval around the prediction. Thus, in an award being supported by statistical evidence, the awarders' judgements at the top end would sensibly be guided more by statistics than the script scrutiny evidence.

the judgemental grades to be spread out. There is no sense in senior examiners having to make judgements at (say) adjacent grades, as the differences between the candidates' achievements at adjacent grade boundaries will be relatively small. Similarly, writing (for example) performance descriptions for adjacent grades would not be tenable. Further, the closer the judgemental grades are to each other the greater the likelihood becomes of 'kinks' (i.e., large changes in the conversion rate) on the Uniform Mark Scale (UMS), which is not desirable.

5. RELIABILITY OF MARKING

If the grade boundaries are closer together because there are more grades (see section 1) the awarding bodies will be even more reliant on their examiners and centres to mark learners' work accurately. Evidence suggests that centres marking coursework (or controlled assessments) for candidates of homogenous (and, particularly, exceptional) ability already find it more difficult to differentiate between the candidates and therefore mark the work less reliably than centres in which the candidates' ability is more heterogeneous (Pinot de Moira, 2005). To promote even more accurate marking, the awarding bodies would need to ensure that mark schemes and any additional guidance on marking, whether for the examining teams or for centres, was even more specific. Further, ensuring that the marks were suitably spread, i.e., promoting good differentiation between the grades, would be even more vital.

6. CREDIBILITY

It is important for any examination to be perceived as valid and accurate, if those who use the results are to do so with confidence. While there is little reason to suggest that a customer's judgement of an examination's validity would be affected by the number of grades used to report the results, that may not be true in the case of reliability. Thinking back to the quote from Willmott and Nutall in section 1, a customer who merely counts the number of times that he or she disagrees with the examination will conclude that reliability has deteriorated if the number of grades is increased. In contrast, a customer who not only notes the number but also the severity of the discrepancies between his or her own judgement and the examination may conclude that the reliability has increased if the number of grades is increased. In short, because customers' assessments of the reliability of examinations are necessarily informal the use of a large(r) number of grades may lead to the reliability of an examination being erroneously judged poor.

7. FAIRNESS IN THE DECISION MAKING SURROUNDING ANY SELECTION PROCESS USING THE GRADES

As Cresswell (1986) pointed out, the use of a small number of grades for reporting examination results causes greater emphasis to be placed upon other selection criteria. If the other selection criteria are less reliable than the examinations, greater reliance upon them will lead to less reliable selection decisions. If the examinations are more reliable than the other criteria in use, then the use of more grades for reporting examination results might, if it caused less recourse to the other selection criteria, increase the reliability of selection decisions. However, reliability is concerned with the consistency with which the same people are chosen on different occasions or by different selectors. Better decisions, in terms of choosing the right people, are likely to result if additional criteria are used alongside examination results since a combination of relevant criteria will have greater predictive validity than would one or two of them. Any tendency, therefore, which the use of many grades to report examination results might cause, to forsake other selection criteria might, ultimately, work against the interests of fairness in decision making.

Further, by increasing the fineness of the distinctions which selectors can make, it imposes a greater responsibility upon them to consider carefully the justifications for the minimum qualifications which they set and therefore may increase the complexity of their task (although whether, in the current context of A level and selection of University applicants, Admissions Officers will see it that way is another matter!).

8. CONCLUSION

In summary, there are a variety of issues to consider when increasing the number of grades in a qualification. The more grades that are used, the more the number of candidates misclassified rises but the more the severity of each misclassification falls, which may outweigh the disadvantage that fewer candidates will receive their actual true grade. Further, increasing the number of grades may require longer test papers, which would increase the burden of the examination. Thought may also need to be given as to which, and how many, grades should be judgemental. Fundamentally, the number of grades should be few enough to reflect the reliability of the mark scale (and the marking), while also being large enough to ensure that the mark scale is not reduced to too few, fairly coarse, categories. Moreover, in order to interpret candidates' grades as being identifiably distinct in subject specific terms, the number of grades depends upon the number of usefully distinct subject specific criteria which can be formulated – which is unlikely to be large[6]. More generally, while increasing the number of grades will probably not diminish the accuracy of selection procedures and may (as may be hoped) enhance it, there are also issues of credibility to be borne in mind. In particular, because customers' assessments of the reliability of examinations are necessarily informal the use of a large number of grades may lead to the reliability of an examination being erroneously judged poor.

The number of grades used to report achievement on any given examination therefore depends on the relative importance of a myriad of coexisting factors. Consequently, if consideration is being put towards extending the number of grades on that examination, it is prudent to ensure that common sense prevails and that fundamental questions such as those below do not become obscured in the fog of technical discussions:

Question 1: Is extending the number of grades necessary, i.e., will it really provide the answer to whatever the perceived problem is, or will it cause more problems than currently exist?

Question 2: If the root of the reason for any proposed extension to the number of grades available in a qualification is to increase the differentiation at the top end, can the qualification support the intense scrutiny being placed on those top end grades?

Question 3: Can an alternative approach be found which may be more appropriate?

Lesley Meyer
Senior Research Associate

[6] Consider how many distinct levels can be explicitly defined within the range of achievements spanned by candidates passing A level, for example – it is perhaps questionable whether this is higher than (or even as many as!) seven.

REFERENCES

Cresswell, M. J. (1984) A level grades - how many should there be? *AQA internal report.*

Cresswell, M. J. (1986) Examination grades: how many should there be? *British Educational Research Journal, Vol 12, No. 1, 1986, pages 37-54.*

Pinot de Moira, A. (2005) Do examiner characteristics affect marking reliability? *AQA internal report.*

Pinot de Moira, A. (2008) Statistical predictions in award meetings: how confident should we be? *AQA internal report.*

Stockford, I. (2011) How accurate can exam grades be? *AQA internal repo rt written for the part y political conferences.*

Wheadon C. & Stockford, I. (2010) Classification accuracy and consistency in GCSE and A level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009. *AQA internal report.*

Willmott, A. S. & Nuttall, D. L. (1975) The Reliability of Examinations at 16+ (Basingstoke, Macmillan).