Monitoring the difficulty of tiered GCSE components using threshold marks for grade C

Vikas Dhawan Research Division

Introduction

GCSE results are reported on a grade scale from A* (highest) to G. Since this scale covers a wide range of attainment, many GCSEs are divided into two tiers, Foundation and Higher. The Foundation tier covers grades G to C and the more difficult Higher tier covers grades D to A*, with grade E often allowed for those candidates who just miss grade D. Centres enter candidates to either the Foundation tier or Higher tier, depending on the centres' judgement of which level is most appropriate for their candidates.

Table 1 shows the grades available on the Foundation and Higher components of a tiered GCSE unit. Example raw grade thresholds – the minimum raw mark required for each grade – are also shown for one session's examinations (both examinations were marked out of 60). The thresholds reflect the difficulty of the examinations. For example, looking at the grade C thresholds for the two example examinations, candidates had to score at least 33 marks on the Foundation examination for a grade C, but only needed 19 marks on the harder Higher examination. The overlapping grades C and D must represent the same standard of attainment on both the tiers so that the resulting award is fair to all candidates. Usually, some items (question sub-parts) are common to both the Foundation and Higher examinations, and candidates' performance on these common items can help inform the grade-thresholds setting process.

Table 1: Grades available in GCSE tiered components and sample grade thresholds

				Overlapping grades					
Higher tier grade	A*	A	B	**C**	**D**	E	*Ungraded*		
Higher tier – minimum raw mark for grade	46	38	28	19	15	13	0		
Foundation tier grade				**C**	**D**	E	F	G	*Ungraded*
Foundation tier – minimum raw mark for grade				33	27	22	17	12	0

The process of setting grade thresholds is known as awarding. Awarding is carried out for each examination (i.e. each session) under the procedure laid down by the GCSE regulator in the Code of Practice (Ofqual, 2011), which states that the purpose of awarding is "to ensure that standards are maintained in each subject examined from year to year". Thresholds are set by expert judgement informed by a wide range of statistical and candidate-performance evidence.

The study reported in this article aimed to explore simple ways of monitoring the relative difficulty of tiered components by considering the difference between the grade C thresholds. If the difficulty of the question papers is as intended, and the grade thresholds have been set correctly, the C threshold mark on the Foundation paper will be higher – as a proportion of the paper total – than the C-threshold mark on the Higher paper. This is because the Higher tier paper will contain a greater proportion of difficult items on which grade C candidates would not be expected to accrue many marks. In fact OCR, Cambridge Assessment's GCSE awarding body, has set itself the demanding target of constructing Foundation and Higher tier question papers such that the C-threshold is at around 85% of the Foundation paper total, and around 40% of the Higher paper total, leading to a target difference between Foundation and Higher tier C-thresholds of 45 percentage points.

There are two basic reasons why the difference between C-thresholds might not be as intended: either one or both question papers might not have been at the target difficulty; or one or both of the C-thresholds might not have been set on the right mark. Table 2 summarises these two reasons in the context of a difference in grade C thresholds having been found to be smaller than designed, or even the wrong way round. Note that the existence of a large difference between the C-thresholds is not a foolproof indicator that all is well. Incorrectly targeted question papers, or incorrectly set grade thresholds, would be just as likely to increase the difference between the Foundation and Higher C-thresholds as to decrease it. Additionally, factors might combine to reduce or increase the difference in thresholds. However, the difference between the C-thresholds is easy to calculate and, when combined with other indicators, might prove useful for routine monitoring of the technical qualities of assessments. Thus in the present study we calculated these differences for two examination sessions, and we present our findings in this article.

Table 2: Potential reasons for a small – or even reverse – difference between the grade C thresholds of tiered examinations

	Foundation tier	*Higher tier*
Test construction	Too difficult at C	Correct difficulty
Awarding	Low C-threshold to compensate	Correct threshold
Test construction	Correct difficulty	Too easy at C
Awarding	Correct threshold	High threshold to compensate
Test construction	Correct difficulty	Correct difficulty
Awarding	Threshold set too low	Correct threshold
Test construction	Correct difficulty	Correct difficulty
Awarding	Correct threshold	Threshold set too high

For a more theoretical discussion on the concepts related to tiered papers see Wheadon and Béguin (2010) and Good and Cresswell (1988a, 1988b, 1988c).

2. Method

The assessments selected for this study were from the GCSE tiered examinations conducted by the awarding body OCR in two sessions, June 2009 and June 2010. For each pair of Foundation and Higher tier components, the C threshold marks were expressed as percentages of the total mark available and the difference between these percentages was calculated (Foundation C-threshold minus Higher C-threshold). Table 3 shows an example of this calculation using one component-pair from the June 2009 session.

Table 3: Example of the calculation of the difference in grade C thresholds between a Foundation (F) and Higher (H) tier examination

C threshold		*Paper total*		*C threshold % of Paper total*		*Difference in % points*
F	H	F	H	F	H	F–H
33	19	60	60	55.0	31.7	**23.3**

3. Results

Figures 1 and 2 summarise the difference in percentage points between the Foundation and Higher tier grade C thresholds from the June 2009 (n=98) and June 2010 (n=115) sessions respectively. The tallest bar in the histogram in Figure 1 shows that, for the June 2009 component-pairs, the Foundation tier C-threshold was between 15 and 25 percentage points above the Higher tier C-threshold for just under 35% of the component pairs. A normal distribution curve is also shown on the graphs for reference. The vertical line at 0.0 on the x-axis shows the point where the C-thresholds were set at the same percentage of raw marks on both tiers; the vertical line at 45.0 percentage points on the x-axis shows OCR's target difference.

The average difference between Foundation and Higher C-thresholds was approximately 20 percentage points in both sessions. Only a small – but apparently increasing – percentage of the component-pairs had C-thresholds which met OCR's target of 45% percentage points between the C-thresholds. This is not in itself particularly alarming, since one purpose of setting targets is to encourage improvements, and there is no suggestion from these data that candidates received incorrect results.

The astute reader will have noticed that three components had 'reversed' C-thresholds, that is, the Foundation tier C-threshold was set at a lower percentage than the Higher tier C-threshold. These three component-pairs were fairly idiosyncratic and are now obsolete. However, further statistical information will now be presented for one of these component pairs as an example of a straightforward investigation that can be done to diagnose possible causes of a reverse – or small – difference in the C-thresholds. Straightforward investigations such as these can be carried out routinely for any component-pair where the threshold-difference is less than desired, and the results used to focus more thorough investigation and improvements.

Figure 3 shows example results of a simple statistical analysis of item *facility* values to investigate a component-pair found to have a smaller than desired difference between the Foundation and Higher C-thresholds, or even a reverse difference, as was the case here. The 'facility' of an item is the average mark scored on the item divided by the maximum for the item. For example, if candidates scored on average 1 mark on an item worth 2 marks, the facility of the item for these candidates is 0.5. Facilities

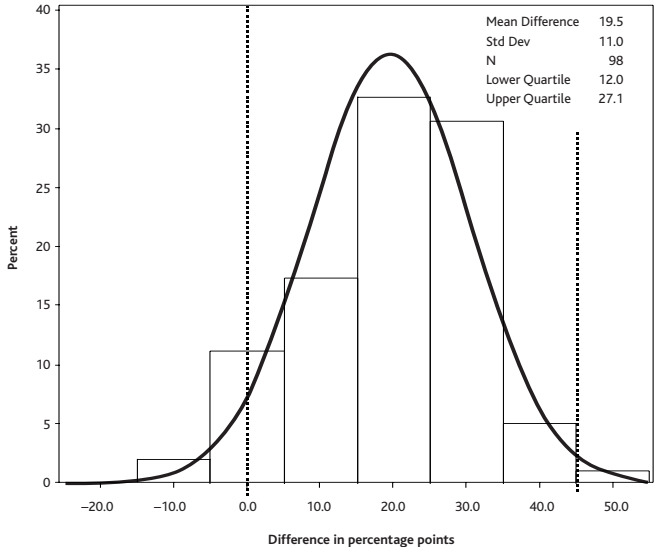

Figure 1: Difference in grade C thresholds (as a percentage of total mark) between the Foundation and the Higher tier components, June 2009

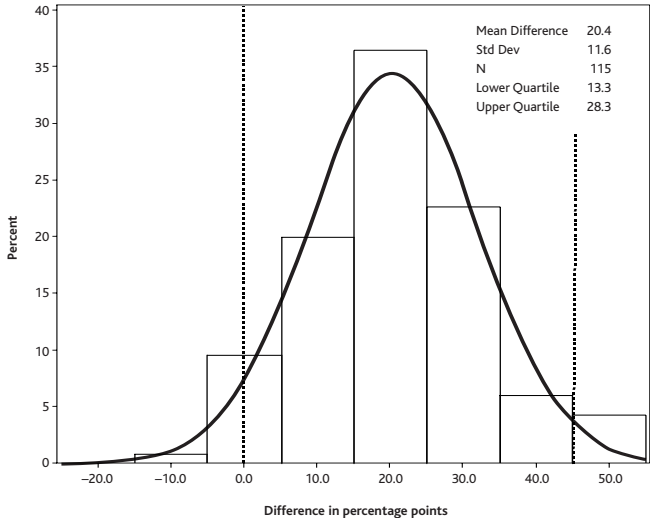

Figure 2: Difference in grade C thresholds (as a percentage of total mark) between the Foundation and the Higher tier components, June 2010

must be between 0 (nobody scored anything) and 1 (everybody scored full marks). Some 10 items, worth a total of 16 marks, were on both the Foundation and Higher tier question papers. These common items are labelled C1, C2, etc., in Figure 3. The remaining items on the Foundation paper are labelled F1–F20, and the remaining items on the Higher paper are labelled H1–H19. Note that we do not expect the facility values of the common items to be the same on both tiers, since we expect the Higher tier to attract a more able entry than the Foundation tier – leading to higher facility values for the common items on the Higher tier.

Two key observations can be made from Figure 3. First, the high facility values for many of the common items show them to be amongst the easiest items on both papers; this is unexpected, since presumably the common items were intended to be pitched at the overlapping grades, and therefore to be amongst the hardest on the Foundation paper and the easiest on the Higher paper. Although the harder common items, C3, C8 and C10, did have lower facility values for Foundation candidates than Higher candidates, the majority of the common items had similar – high – facility values for both. Moreover, on the Foundation paper, the

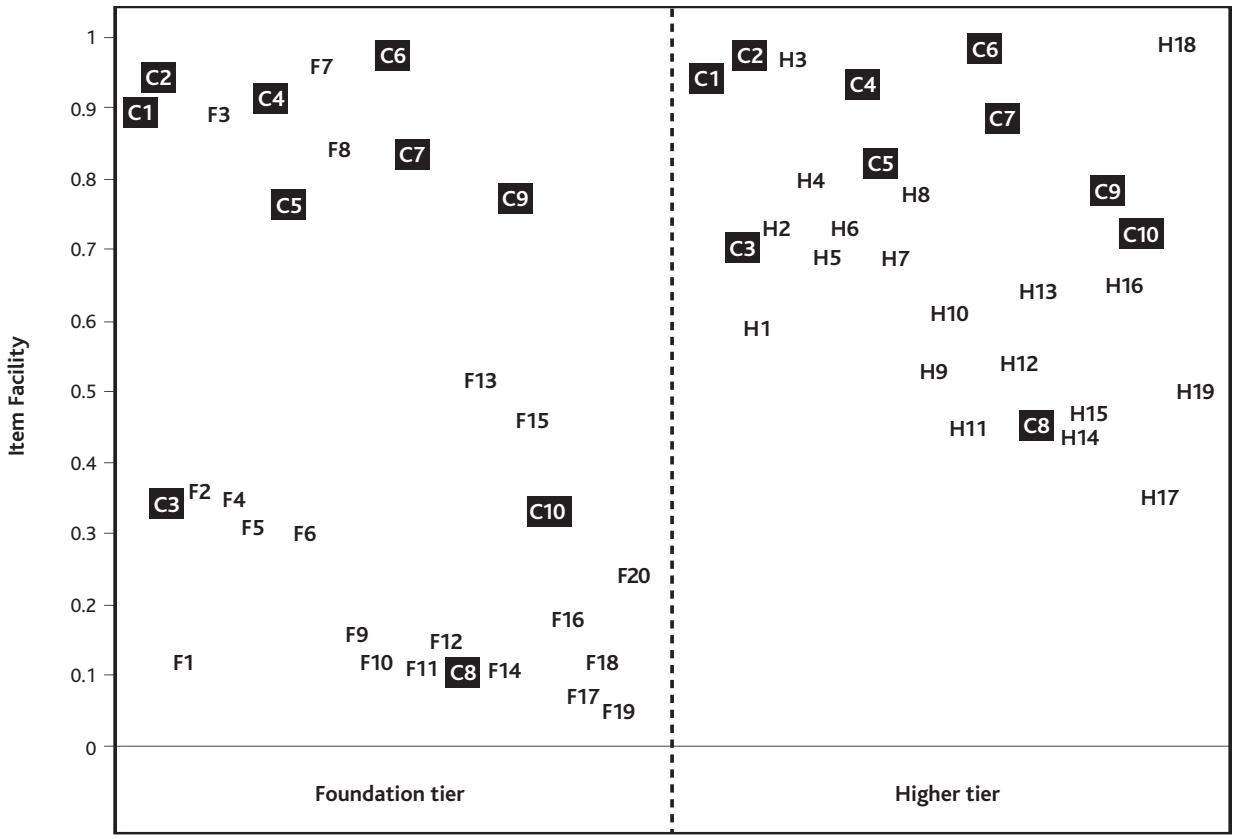

Figure 3: Facility values of items, Foundation and Higher Tier

difference between the facility values of the easy common items and the non-common items was generally greater than the difference on the Higher paper – implying that the non-common Foundation items were generally harder than the non-common Higher items. This is the second key observation from Figure 3: many of the non-common Foundation items were apparently harder than the non-common Higher items. Thus in this example, it appears that issues relating to test construction were the major cause which led to the unexpected difference in the C thresholds. Further investigation – by the teams involved in setting these examinations – would be required to explain what caused the mis-targeting and improve future papers.

4. Discussion

In this article we have looked at a simple indicator for monitoring the difficulty-targeting of examinations in tiered GCSEs, namely the difference in percentage points between Foundation tier and Higher tier Grade C thresholds. It was argued that if the Foundation tier C-threshold was not at a considerably higher percentage mark than the Higher tier C-threshold, at least one of the examinations must have been mistargeted, or at least one of the thresholds must have been set on the wrong mark. We showed that a straightforward comparison of the facility values of items common to both examinations with the facility values of items on only one examination allowed us to diagnose the likely cause of an unexpected difference between C-thresholds. In the case that we examined, the likely main cause was that the common items were too easy and the non-common items on the Foundation examination were too hard. Further investigation, including a detailed review of the items, would be required to identify what caused this mistargeting.

The reason for using grade C threshold comparisons and item facility values was that they are routinely available. The production of graphs such as Figure 1 was easily automated and so can be incorporated into routine quality monitoring, and item facility values are already routinely produced for examinations and so are available for initial investigation of any tiered examinations flagged by the C-threshold comparisons. More sophisticated statistical and psychometric techniques, such as Rasch analysis, are more accurate and powerful, and would more clearly indicate whether the C-thresholds were set at the same standard on both tiers, and the relative difficulty of the non-common items, but at the cost of increased complexity of production. Their use might best be reserved for investigating issues flagged by simpler means such as those used in this article. However, all statistical and psychometric techniques for comparing the difficulty of the examinations and the standard of the C-thresholds, whether simple such as facility values or more sophisticated such as Rasch analysis, depend on the common items for linking the examinations. This leads to several caveats. First, the common items may not fully represent the entire content of the examinations, and it is possible that candidates performed differently on the non-common items. Moreover, by trying to make the items suitable for both tiers, the item writers might have introduced features to the common items that caused candidates to perform differently on these items, or which prevented the items from discriminating well between candidates of different ability. Secondly, measurement error depends in part on the number of items in a test, and so the measurement error associated with candidates' scores on the 'sub-test' comprised of the common items will likely be greater than the measurement error associated with candidates' scores on the whole examination. Kolen and Brennan (2004) recommend that at least 20% of the items on two tests to be equated should be common. However, this is on the assumption that the examinee groups are not very different. In the

context of tiered examinations, we are dealing with different groups (we expect more able candidates to enter the higher tier) and, as Klein and Kolen (1985) (cited in Cook and Petersen, 1987) demonstrated, when examinee groups are different the proportion of items common to the tests become important. Thirdly, the common items are rarely in the same order on both examinations, and this might affect the difficulty of the items. For these reasons, the linking of the tiers via candidates' marks on the common items should be treated with caution.

It should be noted that having many items with low facilities on a Foundation paper, or many items with high facility values on a Higher paper, does not necessarily mean that the papers were mistargeted: candidates might have entered for the wrong tier.

The comparison of grade C thresholds of tiered examinations is not on its own a complete method for identifying issues with difficulty targeting or standard setting. Comparing thresholds at grade D might result in a different interpretation. Moreover, issues with difficulty targeting or standard setting might not be reflected in reduced or reversed differences between the Foundation and Higher thresholds. The method of comparing C grade thresholds is recommended because it is straightforward, easy to automate and can then be done routinely as part of a wider monitoring system.

When a reduced or reversed difference between Foundation tier and Higher tier C-thresholds is detected, it is important to understand what has caused it. If items did not function as intended and an examination was harder or easier than it should have been, it is appropriate to set lower or higher thresholds respectively to compensate. Thus an unexpected difference between Foundation and Higher tier C-thresholds does not imply that either threshold was wrong or that the standards applied were not comparable; it can simply reflect the fact that the difficulty of one of the examinations was not optimal for its tier. Once this has been detected by means of the simple techniques described in this article, further investigations can take place to identify improvements for future examinations.

Acknowledgements

I would like to thank my colleagues Nick Raikes, Tom Bramley, Beth Black and Mike Forster for their advice.

References

Cook, L.L. & Petersen, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*. **11**, 3, 225–244.

Good, F.J. & Cresswell, M.J. (1988a). Grade awarding judgements in differentiated examinations. *British Educational Research Journal*, **14**, 3, 263–80. http://www.tandfonline.com/doi/pdf/10.1080/0141192880140304. Accessed on 22nd February, 2012.

Good, F.J. & Cresswell, M.J. (1988b). *Grading the GCSE*. Secondary Examinations Council: London.

Good, F.J. & Cresswell, M.J. (1988c). Placing candidates who take differentiated papers on a common grade scale. *Educational Research*, **30**, 3, 177–189. http://www.tandfonline.com/doi/pdf/10.1080/0013188880300302. Accessed on 22nd February, 2012.

Klein, L.W., & Kolen, M.J. (1985, April). *Effect of number of common items in common-item equating with non-random groups*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Chicago.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking. Methods and practices*. 2nd edition. New York: Springer-Verlag.

Ofqual (2011). GCSE, GCE, Principal Learning and Project Code of practice. http://www.ofqual.gov.uk/for-awarding-organisations/96-articles/247-code-of-practice-2011. Accessed on 28th February, 2012.

Wheadon, C. & Béguin, A. (2010). Fears for tiers: are candidates being appropriately rewarded for their performance in tiered examinations? *Assessment in Education: Principles, Policy & Practice*, **17**, 3, 287–300. http://www.tandfonline.com/doi/pdf/10.1080/0969594X.2010.496239 . Accessed on 22nd February, 2012.

An investigation on the impact of GCSE modularisation on A level uptake and performance

Carmen L. Vidal Rodeiro Research Division

Background of the study

Over the past few years modular assessment has been gaining popularity in England, particularly in large scale assessments such as the General Certificates of Secondary Education (GCSEs), which are taken by the majority of 14–16 year olds. Instead of being assessed at the end of a two-year course by following a linear syllabus, GCSE modular courses allow the assessment to take place in specified sessions in both the first and second years of the course. When multiple assessment paths exist for the same subject, it is left to individual schools to decide whether the assessment should be modular or whether candidates should enter for a linear examination.

However, it has recently been suggested that these modular assessments led to changes in learning opportunities and in the interaction between learning and assessment. In particular, modular assessment has been criticised for leading to fragmentation of learning and to a lack of coherence in the learning experience, endangering what is called synoptic understanding (Hayward and McNicholl, 2007), as students have little time for reflection, skill development and consolidation of learning. Furthermore, modular assessment might not provide opportunities for deep learning and it might, instead, encourage a climate of cramming (Priestley, 2003). In addition, the increased assessment load can lead children to spend more time revising for the next exam, rather than simply benefiting from learning (Hodgson and