Comparing difficulty of GCSE tiered examinations using common questions

Vikas Dhawan and Frances Wilson Research Division

Introduction

Tiering is a test design followed in the UK for some GCSE examinations whereby it is intended to develop tests at different difficulty levels (and with different available grades). Teachers or schools then decide what the most appropriate tier is for their pupils. In such a *differentiated* assessment the higher proficiency candidates are allocated to the more difficult 'higher tier', whereas those towards the lower end of the proficiency scale are allocated to the easier 'foundation tier'. 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. The overlapping grades in the two tiers, C and D, are intended to represent the same level of performance, irrespective of the tier on which they may be achieved. Table 1 shows the grades available on the foundation and higher tier components of a tiered GCSE unit.

Table 1: Grades available on GCSE tiered components

				Overlapping grades						
Higher tier grade	A*	A	B	**C**	**D**	*E*	Ungraded			
Foundation tier grade				**C**	**D**	E	F		G	Ungraded

The process of setting grade boundaries (the minimum mark required to attain a grade) for each examination for each session is called 'awarding'. It is based on the procedures laid down in the Code of Practice (Ofqual, 2011a) issued by the examination regulator – Office of Qualifications and Examinations Regulation (Ofqual) – which states that the purpose of awarding is "to ensure that standards are maintained in each subject examined from year to year…". Subject matter experts compare candidate scripts (or coursework, if applicable) at different performance levels and judge them to be worthy of specific grades. The grade boundaries are decided by the experts based on the candidate performance and various sources of statistical evidence (such as teachers' forecast grades, performance of the cohort in the previous sessions, etc.). The complete list of potential sources of evidence which can be used for awarding is given in Ofqual (ibid.). The grade boundaries which are decided by the experts by using these sources of evidence are called 'key boundaries'. Not all grade boundaries are obtained by this 'judgemental' process. The rest are calculated arithmetically to lie between the key boundaries. For the GCSE tiered examinations used in this study, the key boundaries are A, C and F (as well as D on the higher tier only[1]).

In tiered examinations, a comparison of the performance at the overlapping grades can be used to maintain standards between the two tiers. Usually the performance at the common grade C (the highest possible grade on the foundation tier) is used to achieve this objective. This is done by developing some items which are common to both question papers. In this study we define 'common items' as those items[2] which have exactly the same structure, format and wording across the two tiers. Usually the mark scheme for common items should be identical across tiers. Here we distinguish between common items and 'similar' items. We define similar items as those items which test the same content across tiers, but use different wording or question structure. The rest of the questions are unique to each paper and have been referred to as 'non-common' questions in this study. It is intended that the non-common questions should on average be easier than the common questions in the foundation tier and should in general be more difficult than the common questions in the higher tier so that the foundation tier is easier than the higher tier. Such a question paper design is intended to provide more support to the foundation tier candidates and more stretch to the higher tier candidates.

If the questions were not functioning as intended, the foundation tier might be more difficult or the higher tier easier than they should be. This essentially is an issue related to test construction and could lead to unexpected differences in the C boundaries between the two tiers. It is normally expected that the C boundary on the foundation tier should be set at a higher proportion of the maximum numeric mark (paper total) than on the corresponding higher tier, because due to the differences in the difficulty of the tiered papers, pupils at this level of achievement (i.e. grade C) should have to get a greater proportion of marks on the foundation tier to attain the same grade than on the relatively more difficult higher tier. If the C boundaries did not function according to this criterion, then it would suggest that the questions on each tier had not been targeted effectively – the foundation tier questions might be too difficult, or higher tier questions too easy. This conclusion would be relevant even if common items were not used. A negative or a very small difference between the C boundaries could indicate that one or both of the question papers might not have been at the target difficulty. Under these circumstances, there is a risk that the grades received by candidates might not be an appropriate reflection of the level of their understanding or proficiency in the subject area. Along with test construction, another reason why the boundaries could be set at the unexpected place is related to awarding. If the grade boundaries were not set appropriately, the difference between the C boundaries could be negative or very small. The use of common questions allows us to investigate further what the 'real' reason might be.

In this study, we investigated the difficulty of the common questions between tiered components to gather evidence of whether the tiered question papers were functioning as expected or not. We used data from the awarding body OCR. We also explored ways in which the analyses could feed into the process of writing questions for tiered examinations and thereby help in improving the current practice of producing such question papers.

1. True for the data used for this study. From June 2012, however, grade D on the GCSE higher tier is now calculated arithmetically.
2. Sub-parts of a question.

For a more theoretical understanding of how tiered examinations work, see Good and Cresswell (1988a, 1988b, 1988c). Wheadon and Béguin (2010) also give a useful discussion on improving standard setting on tiered tests using Item Response Theory (IRT) models.

Current practice in producing common questions and tiered papers

Tiered assessments have been a common feature of GCSE assessment since the introduction of GCSEs in 1988. Towards the end of the 1990s, the number of tiers used to differentiate between candidates was reduced from three to two in most subjects. The number of tiers to be used in GCSE assessments is regulated by Ofqual. In 2004 the Qualifications and Curriculum Authority (QCA) – Ofqual's predecessor – specified that "the assessment arrangements for GCSEs must … include question papers targeted at two tiers of grades, A*-D and C-G, unless subject criteria or the regulatory authorities indicate otherwise" (QCA, 2004, p.27). However, Ofqual's Code of Practice (2011a) does not specify whether GCSE assessment should be tiered or not. Currently, the Subject Criteria, which are published by Ofqual on a subject by subject basis, specify whether a given subject must have tiered examinations. To the best of our knowledge, no formal motivation for the decision to use tiered examinations or not in a subject is available. It is frequently informally suggested that subjects which have tiered examinations are those such as Science or Mathematics, where the questions targeted at the top grades would be inaccessible to less able candidates, and would thus provide a demotivating assessment experience for these candidates. However, it is unclear, for example, why Latin is tiered, but Classical Greek is not (Ofqual 2011b), or why History is tiered in Northern Ireland, but not in England and Wales (Ofqual, 2012a).

Tiered question papers aim to differentiate candidates of different abilities, while still allowing for comparability between the awarded grades where the papers overlap (grades C and D). There are often differences between foundation and higher tier papers with respect to the style and format of tasks. Foundation tier papers frequently use tasks which are more structured, and use less complex vocabulary and sentence construction. However, common items should be suitable for use in both foundation and higher tiers. Within OCR, there are few formal guidelines for setting common items for those subjects which have tiered examination papers. Where such guidelines exist, they do not typically extend beyond specifying the need to target common items at the overlapping grades C and D. For some OCR question papers, for example, 2359 (ICT) or A353 (Classical Civilisation), the common questions are set in a block of questions which differ in format from the questions specific to the foundation tier (typically objective) and higher tier (typically extended answer). Where common questions form a block of questions, they typically occur towards the end of the foundation tier paper, but at the beginning of the higher tier paper, consistent with the more general approach of putting the most difficult exam questions towards the end of a paper.

Despite the general lack of literature relating to the current practice in setting common items, Ofqual does consider the use of common items in some papers. In a review of standards in GCSE English between 2005 and 2009 (Ofqual, 2011c), it was noted that the tasks common to both foundation and higher tiers may provide more scaffolding than is appropriate for the higher tier. In contrast, a review of GCSE Business Studies (QCA, 2005a) commented that where similar questions were used across tiers (in OCR and AQA 2003 examinations), the question format was more similar to that used in the higher tier papers, and was rather demanding for the foundation tier candidates. However, if different guidelines for question formats are set for different tiers, it seems inevitable that the format of common tasks must compromise tier specific guidelines to some extent.

Ofqual also conducted a series of recent reviews of GCSE standards across time in a limited set of subjects (Ofqual, 2012b). A minority of papers in this series comment favourably on the use of common questions. For example, a review of Biology GCSE standards from 1999–2003 (QCA, 2005b) notes that the use of common questions allows comparison of standards for candidates awarded grade C across tiers. Additionally, a review of Chemistry GCSE standards (QCA, 2005c) in 1998 and 2003 recommended that where common or similar questions are used, identical wording should be used to allow direct comparison across tiers, even if this means that foundation tier wording is used on higher tier papers. It is notable that this is contrary to the recommendations for English (Ofqual 2011c), which criticised the use of foundation tier question wording in a higher tier paper.

Method

A list of 81 pairs of foundation-higher tier assessments (referred to as 'component-pairs' in this study) were obtained from the June 2011 session examinations. The C boundary raw mark of each component was calculated as a percentage of its paper total and a difference (foundation – higher) between the percentages of each component-pair was used to select a potential group of component-pairs. A positive difference here would indicate the expected situation – that the C boundary on the foundation tier as a proportion of its paper total was higher than that on the higher tier. On the other hand, a negative (or a very small positive) difference might suggest that issues related to test construction and/or awarding need to be investigated. The difference between the C boundaries in each pair was also compared against the 'target' or expected difference set by OCR at 45 percentage points – 85% of the foundation paper total and 40% of the higher paper total (Dhawan, 2012).

Figure 1 shows the distribution of differences (foundation – higher) in C boundaries as a proportion of paper total between the two tiers for 81 pairs for the June 2011 session. The graph also gives a few summary values of the differences. A vertical line at 0.0 on the x-axis identifies the point where the C boundary between the two tiers was exactly the same. Another vertical line at 45.0 identifies the target difference in C boundaries between the tiers.

Figure 1 shows that the mean difference in C boundaries in the component-pairs was 23.2 percentage points with a standard deviation of 6.9. The median of the differences was 23.8 and the minimum and maximum difference was 8.3 and 42.5 respectively. As is evident from the figure, no negative difference values in the C boundaries were observed. However, there were a few components with a very low difference in the C boundaries which could flag up some possible concerns in the design of the question papers. There were hardly any pairs which were close to the target difference of 45 percentage points. Table 2 gives the summary statistics of C boundaries as a percentage of paper total for all the components. The table suggests the C boundaries at the higher tier were, on average, close to

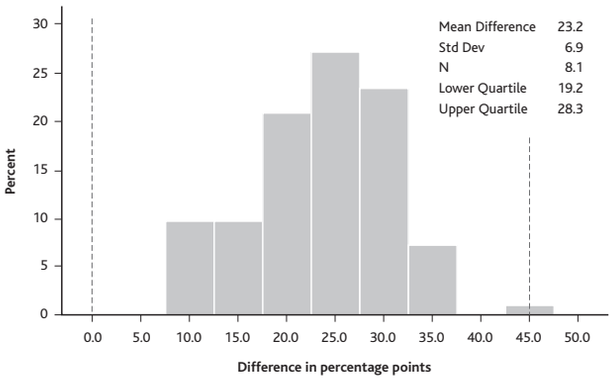

Figure 1: Differences in grade C boundaries (as a percentage of maximum mark) between foundation and higher tiers, June 2011.

the target (40% of raw marks) whereas those at the foundation tier were, on average, lower than the target (85% of raw marks). The lower than targeted C boundaries at the foundation tier indicates why there were hardly any component-pairs near the target difference.

Table 2: Statistics of C boundaries as a percentage of paper total

Tier	**Mean**	StdDev	N	Min	Max	Q1	**Median**	Q3
Foundation	**65.5**	11.1	81	43.6	92.5	57.5	**63.9**	70.4
Higher	**42.3**	12.6	81	18.3	68.0	31.0	**40.0**	52.0

The results given here for the June 2011 session (and those for the June 2009 and June 2010 sessions given in Dhawan, ibid.) suggest that OCR might have set itself a demanding target of achieving a 45 percentage point difference between the C boundaries.

A final list of six component-pairs selected based on the level of difference between the C boundaries of the two tiers and review of question papers and mark schemes is given in Table 3. Note that we have given generic labels to the components. Two component-pairs with the largest unexpected difference, two with the most commonly observed difference between C boundaries and two with a large positive difference were classified respectively as:

a. **Low group** – a difference of less than 17 percentage points;

b. **Median group** – around the average difference of all the pairs; and

c. **High group** – around the target difference of 45 percentage points.

The table also gives the difference between the C boundaries in each pair and the percentage of common items in the paper.

The functioning of items between each component-pair was investigated using Rasch analysis. The Rasch method expresses the estimates of item difficulty and candidate ability on the same scale, called logit or log odds unit scale (Bond and Fox, 2007). This method produces estimates of relative item difficulty which are independent of the ability of the cohort and estimates of candidate ability which are independent of the difficulty of the items. First, we compared the relative ordering of the difficulty of the common items in the tiers. Secondly, we used the common items to equate the two tiers in each pair and compared item difficulty with cohort ability distribution. The common items were used to equate the two tests by applying what is known as the one-step or concurrent method (Hanson and Béguin, 2002; Morrison and Fitzpatrick, 1992). In this method, the student responses from both the tests to be equated are combined in a single dataset and the calibration of the tests is done simultaneously.

Kolen and Brennan (2004, p.271) recommend that the common items should be at least 20% of the length of the total test for equating to be adequate in practice. This is on the assumption that the examinee groups are not very different. However, in the context of tiered exams we are dealing with rather different groups 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 becomes more important". Table 3 shows that the percentage of common items appeared acceptable for all the component pairs.

We then conducted a qualitative review of how common items relate to non-common items within a pair of question papers, and examined how similar questions, which test the same or similar content, varied across the tiers.

Results

Comparing relative difficulty of the common items

The item difficulty values from Rasch analysis were compared for the common items in each pair. If the foundation and higher tiers are assessing the same trait, differing only in overall difficulty, then the common items should have the same *relative* difficulty in both. Data from both the tiers in each pair were analysed separately and the difficulty values of the common items were plotted against each other[3].

3. The separate analyses fix the origin of each scale at the mean item difficulty (i.e. including common and non-common items) on each tier. Therefore the common items will have a different mean difficulty in each tier. The two scales are aligned by 'shifting' the values from one of the tiers by an amount equal to the difference in mean difficulty of the common items (see Wright & Stone, 1979, pp.112–118).

Table 3: Difference in C boundaries as a proportion of paper total between the tiers

Group	*Component-pair label*	*Subject*	*C boundary*		*Paper Total*		*C boundary/ Paper Total %*		*Difference in % pts*	*% of common items*	
			F	*H*	*F*	*H*	*F*	*H*	*F - H*	*F*	*H*
Low	L1	Biology	24	19	55	55	43.6	34.5	9.1	27.6	28.6
Low	L2	Additional Applied Science	24	18	36	36	66.7	50.0	16.7	34.8	38.1
Median	M1	Applied Science	37	23	60	60	61.7	38.3	23.3	41.2	41.2
Median	M2	Physics	31	17	60	60	51.7	28.3	23.3	27.1	28.9
High	H1	Mathematics	35	17	60	60	58.3	28.3	30.0	16.7	19.2
High	H2	ICT	32	11	60	60	53.3	18.3	35.0	50.0	61.5

F=Foundation tier H=Higher tier

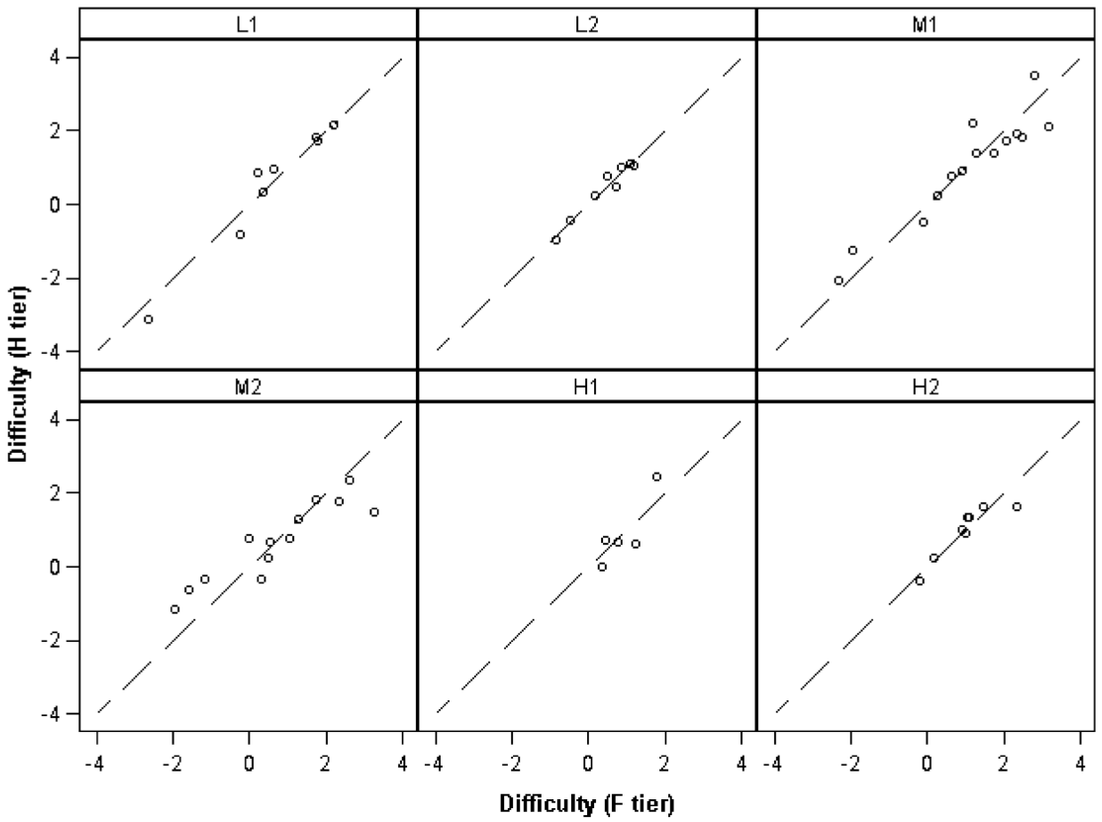

Figure 2: Comparison of Rasch difficulty of common items – Foundation and Higher tier

The results are shown in Figure 2. The x-axis shows the Rasch difficulty on the logit scale in the foundation tier and the corresponding values in the higher tier are given on the y-axis. The items towards the negative end of the scale (-4) indicate easier items, whereas those towards the positive end indicate more difficult items. The line in the middle of each plot is an identity line. Items that fall on this line had the same difficulty values across the tiers. The items below this line were relatively easier on the higher tier, whereas those above the line were relatively easier on the foundation tier.

Figure 2 shows that most of the common items across the six component-pairs were either on or close to the identity line and therefore were of similar difficulty between tiers. There were a few common items, particularly in the Median group (M1 and M2), which did not appear to have similar difficulty and therefore might not have been adequately functioning as common items. From this figure, it appears that the common items in the pairs that were classified into the Low group (L1 and L2) and the High group (H1 and H2) were more or less of similar difficulty. However, it might be due to the fact that the pairs in the Median group had a higher number of common items, some of which did not function as intended.

Overall, it appears that the common items in almost all the pairs had the same relative difficulty on the foundation tier and the higher tier, suggesting that it is reasonable to use common items equating to link scores across the tiers.

Item difficulty and cohort ability

The results from Rasch common item equating for each component-pair are given in Figure 3. The lower part of the graph for each pair shows the estimates of item difficulty after equating. The items towards the left hand side on the x-axis are the easier items and become increasingly difficult towards the right hand side. The items have been identified as common (shown as dots), non-common in the foundation tier (triangles) and non-common in the higher tier (squares). The item estimates are shown here after equating; therefore the common items appear at the same position for both the tiers. The upper part of the graph shows the percentage distribution of ability estimates of pupils on both the tiers. The graph also gives the number of pupils for each tier. Pupils with lower proficiency in this test are shown towards the left hand side of the x-axis and those with higher proficiency are towards the right hand side.

Figure 3 shows that in some components such as L2 and H2 there was hardly any difference in the ability of higher tier and lower tier candidates which suggests that the use of tiering is redundant in these assessments. The figure also shows that the non-common items in some components such as L1 and L2 were very similar in difficulty contrary to the expectation. This effect tends to improve with the increase in the difference in the C boundary between the tiers and the components H1 and H2 have a more clear distinction between the items in the two tiers. H2 gives the best example in this study of the relation between the common and non-common items in which the common items were the easier ones in the higher tier and the more difficult ones in the foundation tier.

The distribution of common items with respect to non-common items is partially dependent on the distribution of items targeted at specific grades. While not all specification grids[4] for the papers analysed in this study give specific grade-targeting information, the specification grids for four of the six papers (H2, M1, M2 and L1) show that common items were indeed targeted at grades C and D, as expected. However, in four of these papers (M1, M2, L1 and L2) the higher tier papers also included

4. A specification grid of a question paper gives a mapping table of items to their target grades.

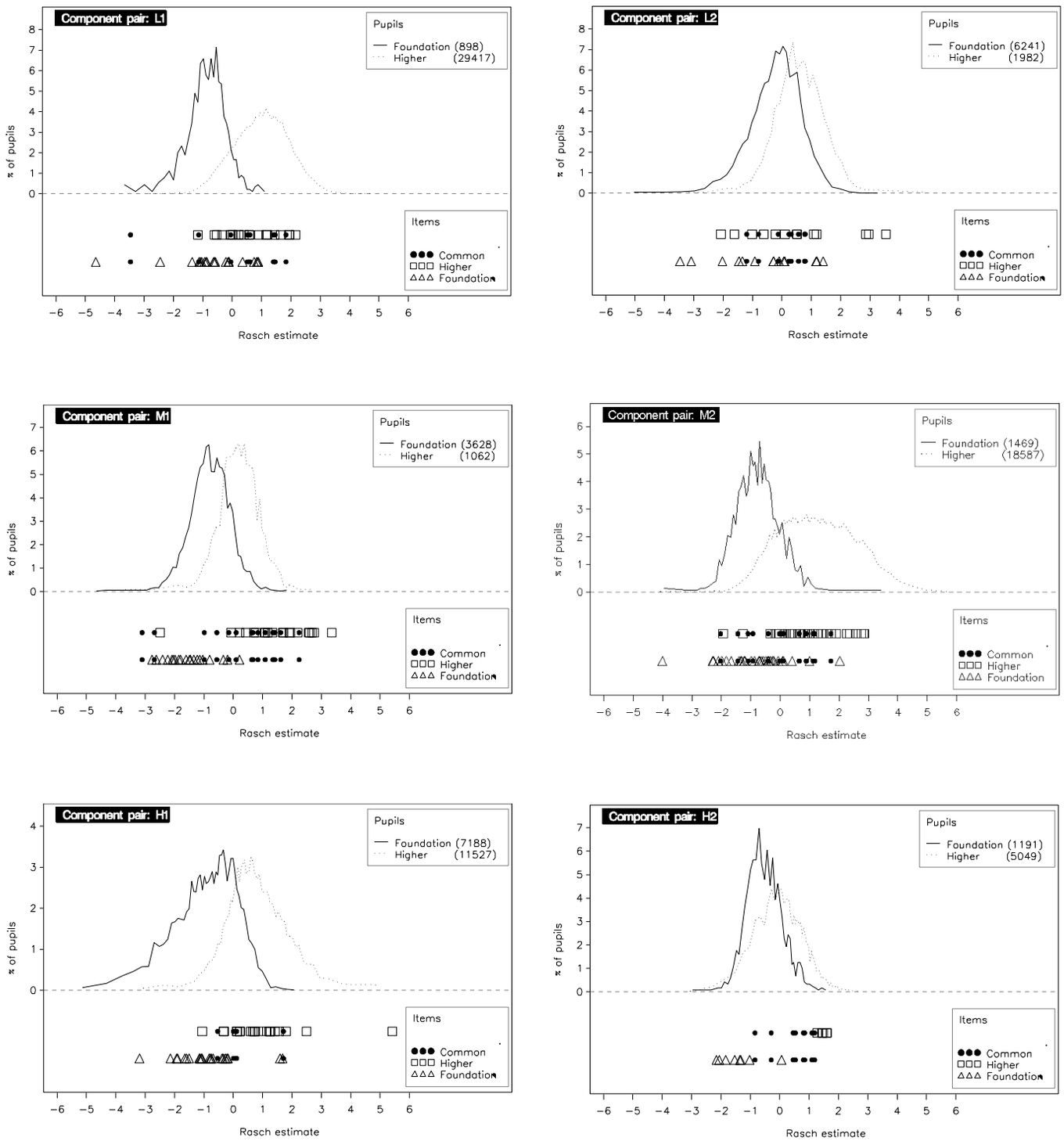

Figure 3: Rasch common item equating, foundation and higher tier

non-common questions which were targeted at grades C and D, whereas the foundation tier paper did not. A closer examination of the non-common items targeted at grades C and D indicates that these items were not as easy as expected for questions at these grade levels in the higher tier. None of the four papers which had non-common items targeted at grades C and D had a high difference between grade C boundaries. For paper H2 (ICT), which had a greater difference between C boundaries across tiers, only common items were targeted at the C and D grades. It seems plausible, therefore, that including non-common items targeted at overlapping grades in the higher tier may have contributed to raising the grade C boundary in the higher tier.

Qualitative review of common items

The question papers surveyed showed different strategies for integrating common and non-common items. Papers H2 (ICT) and M1 (Applied Science), presented common items within a block of questions, which was at the beginning of the higher tier paper and the end of the foundation tier paper, reflecting the fact that the items which are more challenging are typically presented at the end of the paper. H1 (Mathematics) and L1 (Biology) followed a similar pattern, with some variation. Separating common items into one block of questions helps to allow comparability by reducing context effects from other question

parts. However, it is not clear where a block of common questions should be placed in a question paper, given that it is advisable to place more difficult questions towards the end of the paper, and the common items should be either the easiest or the hardest questions depending on tier. It is possible that performance on common questions located towards the end of a foundation tier paper would be lower than if they were located towards the beginning of the paper because candidates had less time to answer them. Kolen and Brennan (2004) and Cook and Petersen (1987) note that common items should not appear in considerably different position on two tests else it might lead to items functioning differently in the two tests.

Rather than using a separate block of questions, M2 (Physics) and L2 (Additional Applied Science) presented common questions in approximately the same position in both tiers, avoiding ordering effects. This might have introduced context effects for the Physics paper, since the common items were often placed towards the end of multi-part questions in the foundation tier, and the beginning of multi-part questions in the higher tier.

Only one of the question papers surveyed (ICT) used several different question types, such as objective questions, short answer and extended answer questions. The ICT paper used constrained objective style questions on the foundation tier paper (with one exception) for non-common items, and used narrative short answer questions worth between two and six marks for the common items and higher tier specific questions. The remaining papers showed no differences in the choice of question type across the two tiers. Although it was tempting to conclude that differentiating the tiers by question type, as exemplified by the ICT paper, has contributed to the target-like patterning of common items across the tiers of this paper, it was difficult to draw firm conclusions on the basis of one paper. Further analysis of a wider range of question papers would be necessary to establish any trends.

We also investigated how similar items, which tested the same or similar content, varied across tiers. The similar items in L2 provided examples of ways in which similar items can be made easier for the foundation tier, despite testing similar content. Both foundation and higher tier items asked candidates to label a picture of a microscope. The wording and layout of the items were identical, except that foundation tier candidates were provided with a list of words from which to choose to label the microscope. This possibly was the reason why the question was much easier at the foundation tier (according to the Rasch estimates).

In the Mathematics paper H1 there was one pair of similar items, in which candidates were shown three scatter graphs and asked to describe the correlation shown in each diagram. The items were differentiated for the tiers by altering the scatter graphs, such that the different types of correlation were stronger for the foundation tier. Although this item was worth three marks on both papers, both foundation and higher tier candidates were asked to describe the correlation shown in each scatter graph. However, to receive full marks for this item, the higher tier had additionally to describe each correlation as strong or weak. Although it would have been possible to use the same layout for the item across tiers, there were differences between the tiers. The higher tier item provided space for candidates to respond immediately below each scatter graph. In contrast, for the foundation tier item each scatter graph was labelled, for example, Diagram 1, and candidates were asked to write their responses further below the scatter graphs, and link their response to the label given to each graph, rather than directly to the graph itself. It seems plausible that adding an additional step of linking responses to a label of a diagram rather than to the diagram itself would require more processing resources, because the label and the link to the actual diagram would need to be retained in working memory. The Rasch estimates for these questions demonstrated that the higher tier question was indeed more challenging, indicating that despite the difference in format between the foundation and higher tier, the difference in content made the higher tier item more difficult.

The qualitative review of items analysed the style, format and content of items in both the foundation and higher tier, with a particular focus on common and similar items. The analysis of individual items suggested that both question style and content play a role in the appropriate targeting of questions. Overall objective style questions seemed to be less challenging, as expected. The distinction between short answer and extended answer questions was less clear, although this may be due to the choice of question style targeted at each tier. For example, the ICT paper (H2) varied the style of questions between the tiers, from objective questions which featured only in the foundation tier, to common short answer questions, and extended answer questions in the higher tier only. However, a more extensive study of more question papers is necessary to determine whether this way of targeting questions to the higher and foundation tiers is effective. Examining questions which were similar, but not identical across tiers aimed to investigate how question structure and layout might contribute to the targeting of questions. However, it was striking that question structure and layout did not always relate to the degree of challenge posed by individual items. This is possibly because the questions investigated were well written and accessible for both tiers, so that the effect of modulation of question style across tiers was minimal. Instead, manipulating the content of similar questions across tiers seems to be of greater importance. This being the case, if questions assess the same content across tiers, it would be advisable to make such questions identical (common) across tiers to allow more effective evaluation of standards between tiers.

Discussion

We found that the grade C boundaries at the higher tier were, on average, close to the target set by OCR (40% of raw marks) whereas those at the foundation tier were, on average, considerably lower (64%) than the target (85% of raw marks). The lower-than-targeted grade C boundaries at the foundation tier explains why few component-pairs in this study were found near the target difference between the C boundaries (45 percentage points). To maintain standards across the tiers, the grade setting procedure should take into consideration the performance on common items. Currently, the emphasis is on maintaining year-on-year standards and the relative performance across tiers might not be given much weight. Where identical common items exist, and can be shown to have the same relative difficulty on each tier, vertical equating outcomes should be taken into consideration when setting common grade boundaries.

The selection of components in this study was based on the assumption that if the C boundaries on the foundation and higher tier were at a similar proportion of the paper total mark, there could potentially be an issue with the test construction (item writing). However, unexpected C boundaries such as this might also be obtained if the grade boundaries were not set appropriately during awarding. Dhawan (2012) presents a number of scenarios where the interaction

between the two issues – test construction and awarding – could lead to unexpected grade C boundaries. For instance, if the foundation tier was comparatively too difficult, it might lead to setting the C boundary very low to compensate. In the current study we focussed on test construction because if the items did not function as intended and an examination was harder or easier than it should have been according to the cohort ability, it would be appropriate to set lower or higher boundaries respectively to compensate. The focus, therefore, was on the review of item writing. In addition, the use of the other overlapping boundary in the two tiers, grade D, might have given slightly different results.

The comparison of the difficulty of the common items might be affected by context effects such as if the items were not in the same order in the two question papers. Ideally, we would want the common items to have the same stimulus and wording, position within a multi-part question, maximum marks and answer space. The mark schemes should also have the same wording and allow the same possible answers in each case. Along with the above criteria, we would expect that the common items were among the most challenging items in the foundation tier and the least challenging in the higher tier. If the tests were not designed keeping in mind these criteria, the consistent functioning of common items across tiers is likely to be adversely affected.

There are some caveats of equating tiered components. The results could be limited by the fact that equating is more appropriate for tests where the cohorts are not too different, whereas tiered examinations are targeted at cohorts expected to be different in ability. The strict set of assumptions for equating results to be adequate recommended by Kolen and Brennan (2004) is unlikely to be fully met in tiered examinations. However, the use of common items for equating is likely to provide more of a robust solution than some of the alternatives.

Large (positive) differences between the grade C boundaries of the two tiers might not be a foolproof indicator that the examinations were functioning as intended. However, comparing C boundaries is a simple procedure which can be carried out in each session. It can be used as an indicator of functioning of tiered components which can be explored further by a qualitative review of the questions. While it is easier to identify items which might not be functioning as intended using statistical evidence, pinpointing the actual cause of the inconsistent functioning could be challenging. Test development is a complex process – one which is influenced by many entities such as the curriculum, the item writers, the awarding bodies and the examinations held in the previous years. Although there are different sources of evidence available, the item writers are still required to 'predict' the difficulty of the items and target them at different grades. Writing of common items is even more challenging because it is expected that the same items should be appropriate in structure and format for both the foundation and higher tiers.

It is worth noting that, if the candidates were not correctly entered in the first place, a comparison of the tiered components is likely to be adversely affected. Future research in this area could focus on the actual process of how the candidates are entered in the tiers, who is involved, which factors are taken into consideration in making this decision, and how the entry decisions vary by different social indicators such as geographical region, gender and school type.

We explored some of the factors that could influence the relative functioning of the tiered examinations using statistical analysis and our perception of why some of the items might not be behaving as intended. Qualitative review of more components, possibly involving some of the item writers and subject experts, might give a better understanding of the functioning of the items. We found that the interpretation of the results was a demanding task because of the paucity of prior literature and specific guidelines – a challenge which the item writers might have to face as well. To conclude, we recommend that:

- a simple procedure such as comparing grade C boundaries could be carried out in each session to identify tiered components which might not be working as intended;
- the functioning of items could be investigated to check if the common items were indeed more difficult than the non-common items in the foundation tier and easier in the higher tier;
- where identical common items exist, and can be shown to have the same relative difficulty on each tier, vertical equating outcomes could be taken into consideration when setting common grade boundaries;
- the statistical evidence can feed into a qualitative analysis of questions to investigate if there were any concerns related to item writing;
- item writers should be provided with a set of specific and written guidelines for writing items in general and tiered examinations in particular;
- Ofqual could publish formal motivation for the decision to use differentiated assessment or not in a subject.

Acknowledgements

We would like to thank the Chairs of Examiners, Qualification Managers and Qualification Leaders of various subjects at OCR for their help and our colleague Tom Bramley for his advice.

References

Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. 2nd Edition. New Jersey: Lawrence Erlbaum Associates, Inc.

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.

Dhawan, V. (2012). Monitoring the difficulty of tiered GCSE components using threshold marks for grade C. *Research Matters: A Cambridge Assessment Publication*, **14**, 18–21.

Good, F.J. & Cresswell, M.J. (1988a). Grade awarding judgements in differentiated examinations. *British Educational Research Journal*, **14**, 3, 263–80. Available online at: http://www.tandfonline.com/doi/pdf/10.1080/0141192880140304. (Accessed 22 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. Available online at: http://www.tandfonline.com/doi/pdf/10.1080/0013188880300302. (Accessed 22 February 2012).

Hanson, B. A. & Béguin, A. A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. *Applied Psychological Measurement*, **26**, 1, 3–24.

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.

Morrison, C. A. & Fitzpatrick, S. J. (1992). Direct and indirect equating: A comparison of four methods using the Rasch model. *Research Bulletin 91–3*. Measurement and Evaluation Center, The University of Texas at Austin.

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

Ofqual (2011b).GCSE Subject Criteria for Classical Subjects. Available online: http://www.ofqual.gov.uk/downloads/category/192-gcse-subject-criteria. (Accessed 18 April 2012).

Ofqual (2011c). Review of Standards in GCSE English 2005 and 2009. Available online at: http://www.ofqual.gov.uk/files/11-09-22-Review-of-Standards-in-GCSE-English.pdf. (Accessed 18 April 2012).

Ofqual (2012a).GCSE Subject Criteria for History. Available online at: http://www.ofqual.gov.uk/downloads/category/192-gcse-subject-criteria. (Accessed 18 April 2012).

Ofqual (2012b). Standards over time. Available online at: http://www.ofqual.gov.uk/standards/92-articles/24-standards-over-time (Accessed 18 April 2012).

Ofqual (2012c). The new GCSE Examinations. Findings from the Monitoring of New Qualifications in French, Business and Geography 2010–11. Available online at: http://www.ofqual.gov.uk/files/2012-03-16-the-new-gcse-examinations.pdf?Itemid=144. (Accessed 18 April 2012).

QCA (2004). The statutory regulation of external qualifications in England, Wales and Northern Ireland. Available online at: http://www.ofqual.gov.uk/files/6944_regulatory_criteria_04(1).pdf. (Accessed 18 April 2012).

QCA (2005a). Review of standards in economics and business studies GCSE and A level 1998 and 2003. Available online at: http://www.ofqual.gov.uk/files/12889_econbusreport.pdf. (Accessed 18 April 2012).

QCA (2005b). Review of standards in biology GCSE 1998 and 2003; A level 1999 and 2003. Available online at: http://www.ofqual.gov.uk/files/12891_biologyreport.pdf. (Accessed 18 April 2012).

QCA (2005c). Review of standards in chemistry GCSE 1998 and 2003; A level 1999 and 2003. Available online at: http://www.ofqual.gov.uk/files/12890_chemistryreport.pdf. (Accessed 18 April 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. Available online at: http://www.tandfonline.com/doi/pdf/10.1080/0969594X.2010.496239. (Accessed 22 February 2012).

Wright, B. D., & Stone, M. (1979). *Best Test Design*. Chicago: MESA Press.

Statistical Reports

The Research Division

The on-going 'Statistics Reports Series' provides statistical summaries of various aspects of the English examination system such as trends in pupil uptake and attainment, qualifications choice, subject combinations and subject provision at school. These reports, produced using national-level examination data, are available on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports.

The most recent additions to this series are:

- Statistics Report Series No.47:
  Uptake of two-subject combinations of the most popular A levels in 2011, by candidate and school characteristics.
- Statistics Report Series No.48:
  A Level Uptake and Results, by Gender 2002–2011.
- Statistics Report Series No.49:
  GCSE Uptake and Results, by Gender 2002–2011.
- Statistics Report Series No.50:
  A Level Uptake and Results, by School Type 2002–2011.
- Statistics Report Series No.51:
  GCSE Uptake and Results, by School Type 2002–2011.