CAMBRIDGE ASSESSMENT

Evaluating the adjacent levels model for differentiated assessment

Tom Bramley

Paper presented at the AEA-Europe annual conference
Tallinn, Estonia, 5-8 November 2014.

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

Bramley.T@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

Abstract

The aim of this study was to compare models of assessment structure for achieving differentiation between examinees of different levels of attainment in the GCSE in England. GCSEs are high-stakes curriculum-based public examinations taken by 16 year olds at the end of compulsory schooling. The context for the work was an intense period of debate among politicians, regulator, examination boards, teachers and public about the appropriate content and assessment structure for the reformed GCSEs announced by the coalition government in 2012.

The focus was on the 'adjacent levels' model for differentiation. In this model, papers are targeted at three specific non-overlapping ranges of grades. The majority of examinees enter for a pair of papers at adjacent levels and receive the highest grade achieved. There is no aggregation of marks across papers. In contrast, in the standard GCSE tiered model examinees take a pair of papers either at foundation tier or at higher tier. The range of grades available overlaps, with C,D and E available on both tiers. A similar tiering model could be applied to the structure with three adjacent papers described above, if the marks on each paper were added together and grade boundaries applied to the aggregate scale, with overlapping grades available as in the standard tiered model.

The adjacent levels model has two important differences from the tiered models: i) there is only one route to each grade; and ii) there is no compensation across papers – the results on one paper are effectively discounted in grading. A review of tiering methods by Baird et al. (2001) discussed the strengths and weaknesses of the model and concluded it was worthy of further consideration.

This study used simulation to compare the three models of differentiation described above in terms of: i) suitability of grade boundary locations; ii) raw score distributions; and iii) reliability (in particular classification accuracy). The simulations were based on an item bank constructed from a mathematics GCSE taken in June 2012. Three papers at adjacent levels of difficulty were constructed by selecting easy, medium and difficult questions from the bank.

The results showed that using an adjacent levels model instead of a tiered model would be likely to reduce reliability in psychometric terms. However, two aspects of validity could increase. The lowest grade boundaries on each paper in the adjacent levels model were at a higher proportion of the paper total than those on the standard model, implying that knowledge of the content of the question paper would give a better indication of what examinees with a given grade would know and be able to do. Furthermore, the single route to each grade removes the ambiguity about overlapping grades inherent in the current system – whether they imply reasonable achievement on a restricted domain of knowledge/skills, or lesser achievement on the full domain.

The discussion considers further the various trade-offs that must be made in choosing a model for differentiated assessment that best serves the interests of the different stakeholders in the system.

Introduction

A variety of models for assessing the wide range of abilities found in the GCSE cohort have been discussed in the past, each with its own strengths and weaknesses. For a detailed discussion of the issues, see Baird, Fearnley, Fowles, Jones, Morfidi & White (2001). They considered four different approaches to differentiation, and evaluated them in the light of the following characteristics that contribute(d) to the GCSE's 'fitness for purpose':

- *Recognising and rewarding positive achievement* – no grades should be the result of an overwhelming failure in the examination (i.e. failure to gain a reasonable percentage of marks);
- *Validity* – the grades should convey a common and generally understood meaning in terms of the inferences they support about what examinees know, understand and can do;
- *Reliability* – if the assessment process were repeated, similar outcomes should be obtained;
- *Comparability of standards* – a grade must represent the same standard of attainment regardless of how it is obtained (i.e. in a tiered structure with overlapping grades, common grades must support the same inferences about what examinees know, understand and can do;
- *Manageability and cost* – different assessment models have different implications for the amount of assessment time, number of papers required etc.

Three of the four models they considered have been widely used at GCSE (and IGCSE):
i) Common papers (all examinees do the same papers and the differentiation is achieved through the mark scheme. This has been used, for example, in History and Art);
ii) Core + extension (all examinees take a core paper with access to less than the full range of grades, and more able examinees can also choose to take an extension paper giving access to the highest grades);
iii) Tiered papers (each tier targets a specific range of ability and has a restricted grade range available to examinees entered for it. There are many varieties of model based on this idea).

The final model considered by Baird et al. (ibid) was at the time the preferred model in Scotland for the Standard Grade exam – the 'adjacent levels' model. Three levels of assessment were produced, each with two grades available and no overlap in these grades. Most examinees would enter for a pair of adjacent levels (the most and least able would just enter for the highest and lowest levels respectively). The grade awarded would be the highest achieved, with no aggregation of marks or 'compensation' across levels. Statistical scaling based on the common examinees in pairs of adjacent levels was used to ensure that the lowest boundary on a higher level was at a higher standard than the highest boundary on a lower level. Baird et al. identified some attractive features of this model:

- The single route to a grade removes the problem of achieving comparability between different routes (present in the tiered model), and the problem of resolving two grades obtained via different routes (present in the core + extension model);
- The single route to a grade means that the papers can be designed to reflect more closely the grade descriptors, enhancing validity and communicability of what examinees with different grades have achieved;
- Most examinees enter for pairs of adjacent levels, reducing the chance of a ceiling effect 'capping' their result or a floor effect resulting in their being ungraded;
- The targeting of content and skills to specific levels gives examinees from across the ability range equal chance to demonstrate their ability and uses examination time more efficiently;
- The between-level statistical equating can be based on large numbers because most examinees take two levels.

The main disadvantage is the lack of compensation – for examinees taking two levels, any achievement on the level for which their grade was ultimately *not* awarded does not count. For example, one examinee may have achieved a grade 1 (the highest grade) and another a grade 2 on the highest level assessment, but their relative performance on the level below might have been such that if the scores were added together the second examinee would have obtained a greater overall total score than the first examinee. In this model, the scores on the lower level would have no bearing on the grades received, since both examinees had performed well enough on the

highest level to achieve a grade on it. As well as having the potential for creating perceived injustices, this 'discounting' of information reduces psychometric reliability. In their evaluation of the strengths and weaknesses of different models, Baird et al. seemed to favour the adjacent levels model, recommending that (along with the core + extension model) it "should be given serious consideration as a possible future model in the GCSE".

The adjacent levels model was used in an OCR Pilot maths syllabus which ran from 2003 to 2007. A report by Stobart, Bibby & Goldstein (2005) compared it with what became the standard 2-tier model and with the existing previous 3-tier model. They found that students tended to prefer the standard 2-tier model to the adjacent levels model – the wider range of difficulties made papers seem more accessible, and they did not have to sit a really hard paper (which was demoralising). Teachers who had taught both were more evenly divided in opinion, feeling that the greater 'stretch' possible with the adjacent levels model could advantage the most able students. There was good agreement (95%) in grade outcome for results created using the adjacent levels aggregation rules and those created using a compensatory approach (putting the boundaries in the place that gave the same overall grade distribution). Stobart et al. deemed the relation between the target grade of the questions (difficulty as intended by the question setters) and their subsequent facility values to be poor, and concluded that normal tiering makes less demand on question paper setters in terms of targeting. They noted that both models had much in common and their main conclusion was that which model was preferable depended on the purpose GCSE mathematics should serve, and what inferences could be made from a grade. The strength of the standard 2-tier model was its accessibility; that of the adjacent levels model its challenge for the more able students. Further analysis of data from this adjacent levels Pilot maths syllabus can be found in Wilson (2013), who found little evidence of any problems caused by lack of compensation for examinees who achieved their grade based on the highest paper (who had almost invariably achieved the highest grade on the middle paper), but evidence that a few examinees who achieved their grade based on the lowest paper had gained more marks than might be expected on the middle paper and thus were disadvantaged by the lack of compensation.

Since the Baird et al. report, one significant change in the examination system has been the large amount of item level data routinely available to analysts as a result of on-screen marking. This makes it possible to carry out more detailed and realistic simulations of different models of differentiation, based on better knowledge of how examinees have performed in the live situation. In the research reported here, a version of the 'adjacent levels' model was explored by simulation for GCSE mathematics. The simulations were based on a linear tiered GCSE Mathematics examination taken in June 2012. The main aim was to compare the simulated outcomes from the adjacent levels model with the outcomes from the 2-tier 2-level model actually used and a 2-tier 3-level model, in terms of: i) suitability of grade boundary locations; ii) raw score distributions; and iii) reliability (in particular classification accuracy).

In the following sections, the structure of the maths GCSE is described, followed by the simulation procedure. The simulated data is then evaluated for how realistic it is, before the results of comparing the different models are presented.

The GCSE Mathematics assessment

This was a tiered examination with a simple structure: foundation tier examinees took Paper 1 and Paper 2 (worth 100 marks in total) and higher tier examinees took Paper 3 and Paper 4 (also worth 100 marks in total). As is standard for tiered GCSE examinations, grades A* to E were available on the Higher tier and grades C to G on the foundation tier. The aggregate grade boundaries and grade distributions are presented in Table 1 below and Figure 1 on the next page.

Table 1: Grade boundaries and grade distributions in June 2012.

	Foundation tier (N=19,048)			Higher tier (N=9,891)			All (N=28,939)	
Grade	Boundary	% cands	cum. %	Boundary	% cands	cum. %	% cands	cum. %
A*	-	-	-	157	14.0	14.0	4.8	4.8
A	-	-	-	125	19.7	33.7	6.7	11.5
B	-	-	-	93	26.3	60.0	9.0	20.5
C	110	36.4	36.4	62	26.0	85.9	32.9	53.4
D	89	21.3	57.8	31	12.2	98.1	18.2	71.6
E	68	17.0	74.7	15	1.4	99.6	11.7	83.2
F	47	13.6	88.4	-	-	-	9.0	92.1
G	26	8.9	97.3	-	-	-	5.9	98.0
U	0	2.7	100.0	0		100.0	2.0	100.0

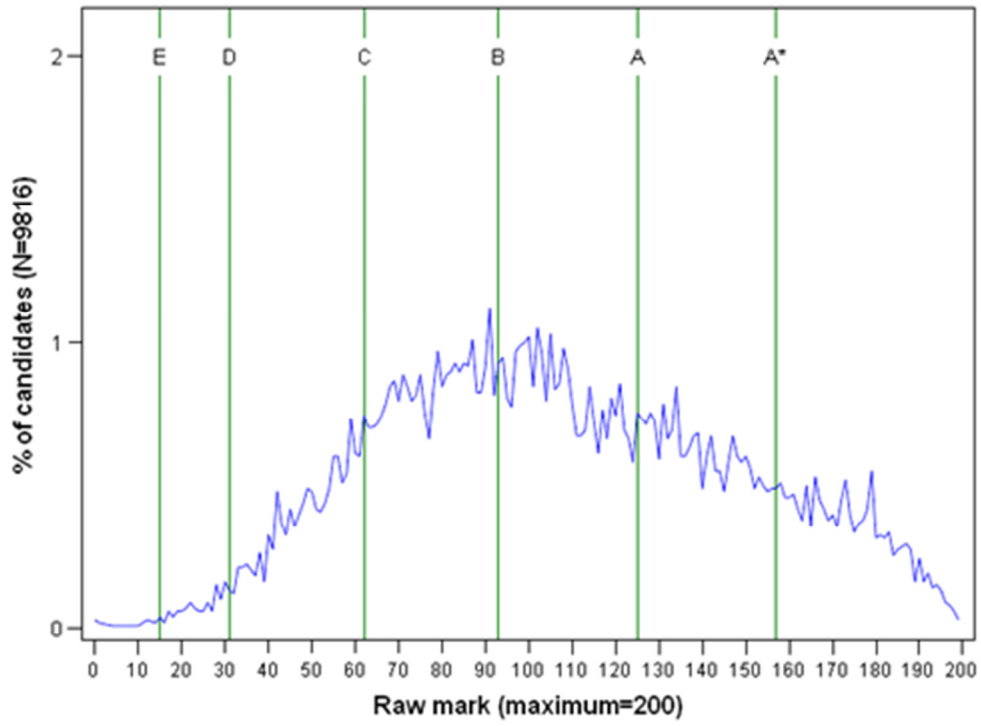

Figure 1a: Distribution of total scores and grade boundaries on the higher tier[1].

[1] Numbers in Figure 1a and 1b do not quite match Table 1 because they were drawn from different data sources. Figure 1 data only included candidates with valid scores for both papers within a tier. Some candidates receiving a grade may have missed one of the papers and had an 'assessed grade'.

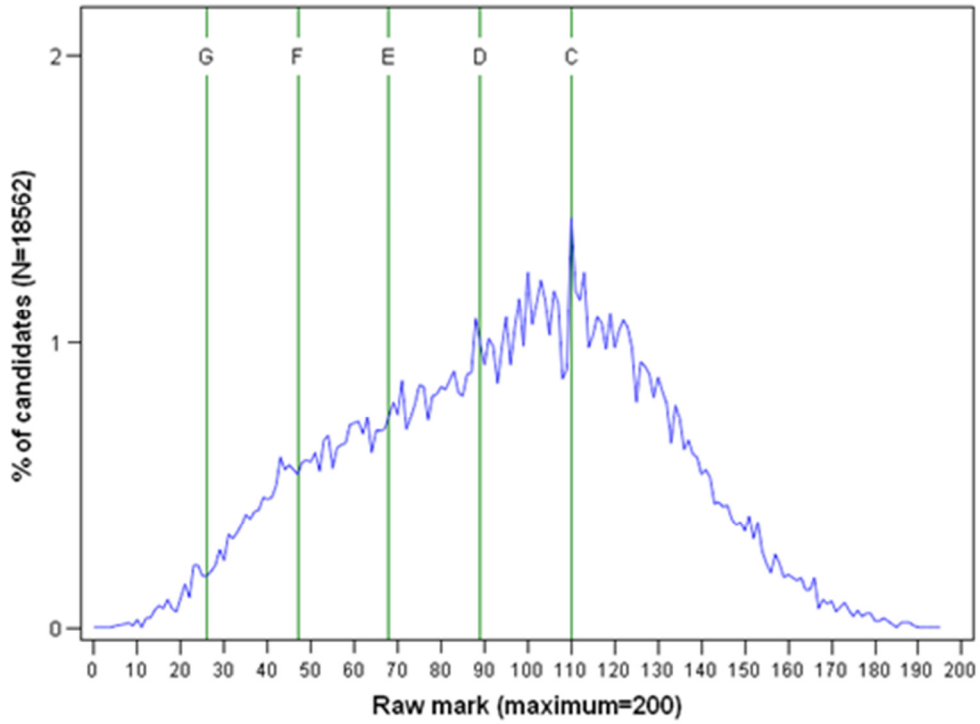

Figure 1b: Distribution of total scores and grade boundaries on the foundation tier.

Two observations can be made from the graphs in Figure 1. First, the higher tier appears better targeted to its 'intended cohort' than the foundation tier because the grade boundaries were well spread across the mark range, whereas on the foundation tier the highest boundary (C) was only at 55% of the maximum, and the boundaries were closer together[2]. Secondly, the boundaries of the overlapping grades on the higher tier (C,D and E) were all at relatively low percentages of the maximum mark, as were the lowest grades (G, F and E) on the foundation tier. It should be noted that this was the first examination session for this syllabus so the outcomes in Figure 1 may not be representative of future sessions (or of tiered assessments more generally). But in the existing data for this assessment there is room for improvement in terms of some of the evaluation criteria identified by Baird et al. (2001), in particular recognising and rewarding positive achievement (since several grades were obtainable by low achievement), and validity (on a related point – inferences based on low scores are likely to be less valid because there is more evidence about what the examinees cannot do than what they can do).

An alternative model for differentiation

It has been widely recognised from the outset of the GCSE that some form of differentiation is required for maths. As discussed above, there may be some merit in a model where there is only a single route to each grade, and relatively higher grade boundaries. The model explored here had the following structure:

Paper 1: 60 marks. Grades available – G, F and E.
Paper 2: 100 marks. Grades available – D, C and B.
Paper 3: 60 marks. Grades available – A and A*.

In this model examinees could enter for Paper 1 only, Paper 1 and Paper 2, Paper 2 only, or Paper 2 and Paper 3. The grading rules applied were as follows:

[2] Both papers appear well targeted to their actual cohort, because the peak of the distribution of scores is roughly in the centre of the mark range on both tiers, and the scores are well spread out.

Rule 1: Grades A / A* are achieved by examinees scoring at or above the B boundary on Paper 2, and at or above the A / A* boundary on Paper 3.
Rule 2: Grades D, C and B are achieved by examinees scoring at or above these boundaries on Paper 2. Some of the grade B examinees may achieve a higher grade based on their Paper 3 performance (see rule 1).
Rule 3: Grades G, F and E are achieved by examinees scoring at or above these boundaries on Paper 1 and below the D boundary on Paper 2.

Note that this model is slightly different from the Scottish adjacent levels model described in the introduction in that there is a 'hurdle' in place. Rather than being discounted entirely, the scores on Paper 2 are important for examinees aiming for an A or A* in that they have to get above the B boundary to qualify for the A or A* on Paper 3. This should mitigate to some extent the drawback of loss of information described above.

Note also that there is less assessment in this model than in the tiered model – Papers 1 and 3 are only worth 60 marks making the total possible assessment 160 marks (as opposed to 200 marks). With the better targeting made possible in principle by having papers at three different levels it was hypothesised that similar or better reliability might be achieved with fewer items per examinee.

Details of the simulation

Creating an item bank

The Rasch model was used to calibrate the items from Foundation and Higher tiers onto the same scale. The non-iterative approximate PROX algorithm (Wright & Stone, 1979) which does not require specialist software, was used. The procedure below was followed to create the scale:
Step 1: Each paper (P1, P2, P3 and P4) was calibrated separately.
Step 2: It was assumed that the grade boundaries within a tier had been set 'correctly' in the sense of corresponding to the same ability on each paper (component). Hence the logit shifts necessary to 'horizontally equate' P2 to P1 (Foundation tier), and P4 to P3 (Higher tier) were found[3].
Step 3: It was assumed that the common C and D boundaries[4] had been set correctly in the sense that they corresponded to the same ability on Foundation and Higher tiers. Hence the logit shifts necessary to 'vertically equate' P3 to P1 and P4 to P1 were found[5].
Step 4: The logit shifts were applied to the item calibrations from P2, P3 and P4 to bring everything onto the P1 scale, thus creating an item bank for use in the simulation.
Step 5: Approximate grade boundary locations on the new logit scale were identified – henceforth referred to as the 'bank boundaries'.

There were in fact 18 common items across the tiers, worth 43 marks in total. An alternative linking method based on these common items was not able to satisfactorily reproduce the grade boundaries – implying that the common items had not functioned as common items for the purposes of grading. The calibrations from the higher tier only were used for these common items in the final calibrated bank.

The coherence of the links was evaluated by checking that the logit shift obtained by indirect equating of one component to another via an intermediate component was approximately the same as that obtained by direct equating. This is illustrated in Figure 2 below. It can be seen that the links did form a reasonably coherent set.

[3] As the average of the shifts required to align G,F,E, and D on the foundation tier, and the average of the shifts required to align D,C,B and A on the Higher tier. (Note that A* is not set at component level).
[4] The 'E' boundary was not used as it is only used as a 'safety net' on the higher tier to prevent candidates who just fail to achieve a grade D from being ungraded. It is not expected to be strictly comparable to the grade E on the foundation tier.
[5] As the average of the shifts required to align D and C across the two tiers.

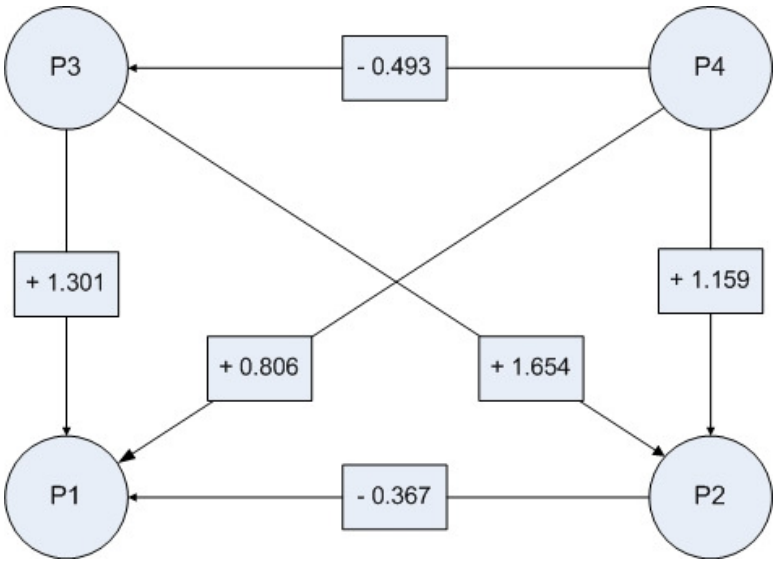

Figure 2: Logit shifts required to align the four components.

Simulating scores

Using the calibrated item bank, item level data was simulated according to the Rasch Partial Credit model (Masters, 1982) for 27,500 examinees (18,000 foundation tier and 9,500 higher tier). The aim was for the simulation to be as realistic as possible, so the approximate ability distribution at each tier was found, based on the average of the equated raw score to scale score conversions (also known as test characteristic curves or TCCs) across both components within a tier. The mean, standard deviation (SD), skew and kurtosis of the actual ability distribution at each tier were used as a basis for generating an ability distribution for the simulated examinees at each tier[6]. Although a distinction was made between higher and foundation tier simulated examinees in terms of generating distributions, scores were generated for all simulated examinees on every item in the bank.

Table 2 below shows that the simulated and actual data had a similar distribution of scores, and Figure 3 shows that even at item level there was a very good relationship between the actual and simulated facility values. (The exceptions are the facility values for some of the common items on the foundation tier, but it was noted previously that these were not treated as common items and the higher tier values only were used in the item bank).

Table 2: Comparison of simulated with actual scores.

	Simulated			Actual		
	N	Mean	SD	N	Mean	SD
Foundation tier	18,000	96.6	37.1	19,048	93.6	35.7
Higher tier	9,500	107.3	42.2	9,891	106.5	41.0

[6] Simulating modified normal distributions using some trial-and-error with the values given in Fleishman (1978) to generate distributions with the desired skew and kurtosis.

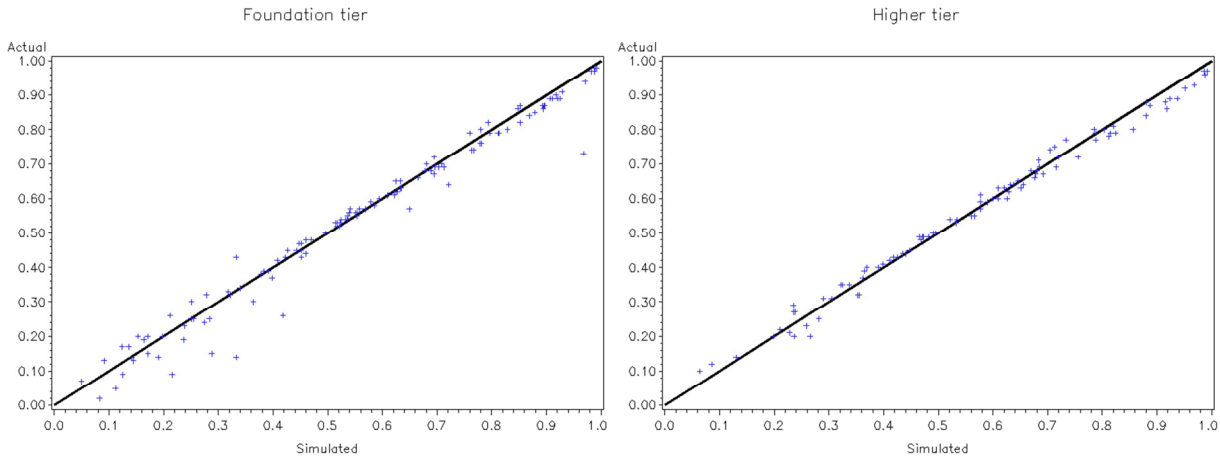

Figure 3: Comparison of simulated with actual facility values. (The reference line is an identity line, not a best-fit line).

Constructing tests for the 'adjacent levels' model

The final item bank contained 72 whole questions comprising 196 part-questions[7] worth a total of 357 marks. The adjacent levels model described previously required the construction of three tests with no common items: a 60-mark easy test, a 100-mark intermediate test, and a 60-mark hard test. This was done by sorting the item bank into order of whole-question difficulty and selecting whole questions from the top and then from the bottom[8] until the 60-mark easy and hard tests had been created. Then the 100-mark test was selected at random from the remaining whole questions. It seemed more realistic and reasonable to select whole questions (rather than part-questions) because tests constructed of part-questions would probably make no sense in reality, whereas it would in principle be possible to assemble papers corresponding to the ones simulated here (although of course they would not necessarily meet requirements for balance of topics etc). Selecting whole questions also meant that there would be the occasional more difficult item (part-question) on the easy paper and easy item on the hard paper.

As would be expected, nearly all the items (56 of 60 marks) on the easy paper were from questions that had originally appeared on the foundation tier; and likewise 58 of 60 marks on the hard paper were from questions that had originally appeared on the higher tier. The intermediate paper had 43 marks from foundation tier questions and 57 marks from higher tier questions.

TCCs for the three tests were created in order to locate the relevant grade boundaries on the raw mark scale of each test corresponding to the boundaries on the bank scale. That is, grades G, F and E on Paper 1; grades D, C and B on Paper 2; and grades A and A* on Paper 3.

Scores of the simulated examinees on each of these tests were then created by adding up the simulated item scores on each of the relevant sets of items. Each simulated examinee was given a grade on each test according to the grade boundaries on the bank scale, and then a grade on the assessment overall according to the three aggregation rules described previously. For simplicity it was assumed that all the simulated foundation tier examinees would take Paper 1 and Paper 2; and that all the simulated higher tier examinees would take Paper 2 and Paper 3 (although the model as stated would allow examinees to choose to take only Paper 1 or only Paper 2).

[7] Some of the foundation tier questions had sub-parts removed because they were also on the higher tier.
[8] In fact an algorithm was written to find the final question(s) once the cumulative mark total was within reach of the required maximum, to ensure the 60-mark total was achieved. (Selecting the top N questions does not guarantee a given paper total mark).

Results

Score distributions and grade boundary locations

Figure 4 below shows the distribution of simulated scores and grade boundaries on the easy and hard papers. The easy paper was only taken by the simulated foundation tier examinees and the hard paper was only taken by the simulated higher tier examinees. Because these papers comprised items almost entirely from the foundation and higher tier respectively, it was possible to compare these distributions with the actual distributions of scores on these items obtained by foundation and higher tier examinees – these are shown for comparison in Figure 5.

It is clear that the simulation accurately represented the actual score distributions, where checking was possible.

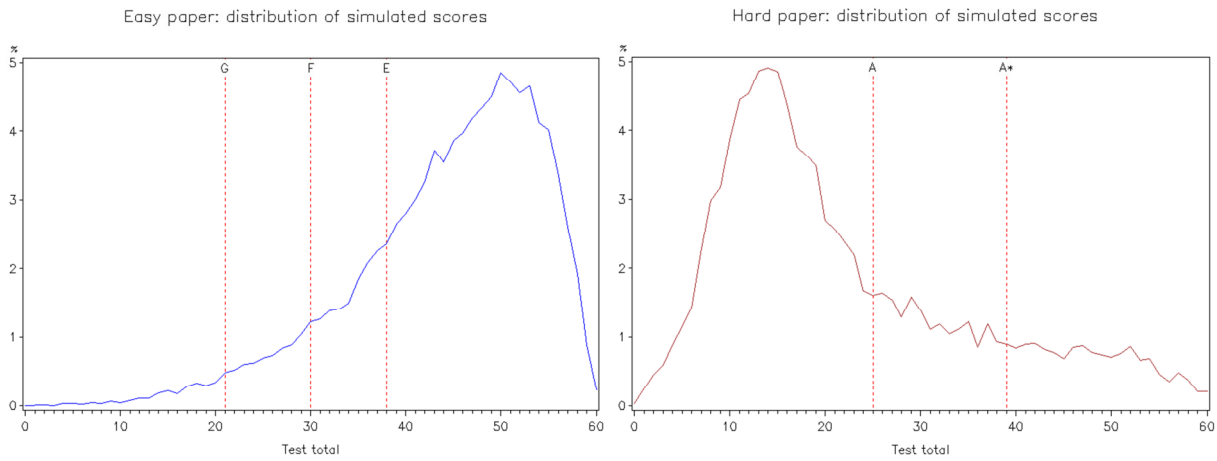

Figure 4: Grade boundaries and score distributions on the simulated 60-mark easy and hard papers.

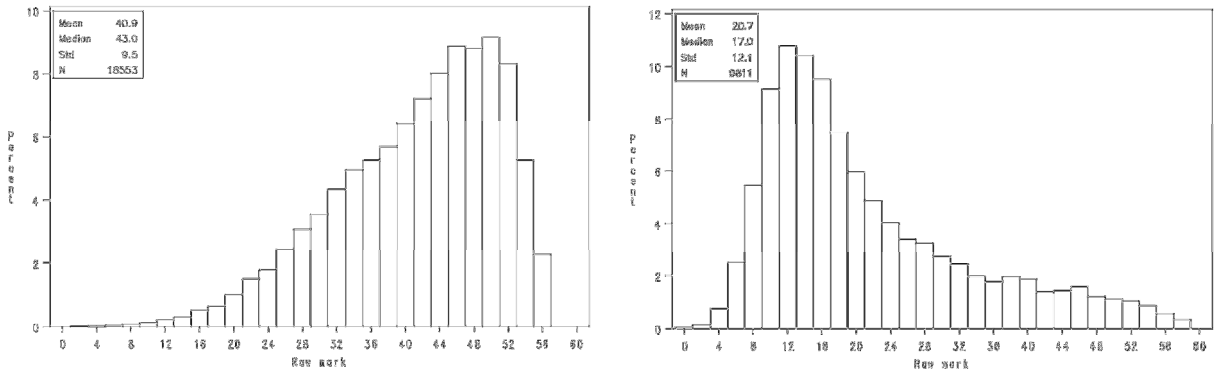

Figure 5: Score distributions on the corresponding actual 56 and 58-mark sets of items.

Figure 6 on the next page shows the distribution of simulated scores for the intermediate ('main') paper, first for all simulated examinees and then broken down by tier. (Note that the percentages on the by-tier graph are relative to the number of simulated examinees in each tier – because there were roughly twice as many simulated examinees at the foundation tier the percentages on the overall graph are not a simple average of the two percentages on the by-tier graph).

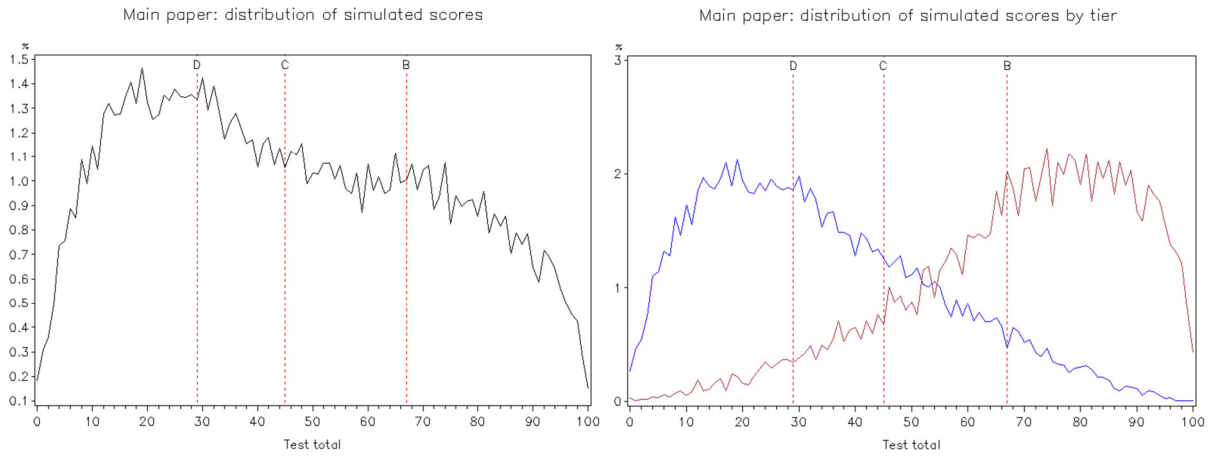

Figure 6: Grade boundaries and score distributions on the simulated 100-mark intermediate (main) paper.

Table 3 shows a comparison of the grade boundary locations in this adjacent levels model with the grade boundary locations in the tiered model. The 'bank scale' grade boundaries were used for the tiered model instead of the actual boundaries (shown in Table 1), to give a fairer comparison.

Table 3: Grade boundary locations on tests constructed from the item bank according to two different models for differentiated assessment.

		Max	Lowest boundary	Lowest as % of max	Highest boundary	Highest as % of max
Tiered model	Foundation	200	29 (G)	14.5%	121 (C)	60.5%
	Higher	200	38 (D)	19.0%	159 (A*)	79.5%
Adjacent levels model	Easy	60	21 (G)	35.0%	38 (E)	63.3%
	Main	100	29 (D)	29.0%	67 (B)	67.0%
	Hard	60	25 (A)	41.7%	39 (A*)	65.0%

It is clear that the adjacent levels model was more effective at achieving the goal of requiring evidence of positive achievement, in the sense of the lowest boundary being at a significant proportion of the total mark. It would presumably help with the validity of inferences about what examinees know and can do, because the question papers themselves (with their more limited range of item difficulties) would better exemplify the kind of questions that examinees could answer. For example, it would presumably be clearer to a stakeholder what a borderline G examinee with a score of 21 out of 60 could do by looking at an 'Easy' paper than it would be what a borderline G examinee with a score of 29 out of 200 could do by looking at a Foundation Tier pair of papers.

An unusual feature of the cohort for this examination was that it had twice as many foundation tier as higher tier examinees. This meant that the mode of the simulated distribution for the (common) main paper was below the lowest grade boundary, which shows that for a significant proportion of examinees the scores on this paper would not have counted towards their final grade. However, although not simulated here, this model would also allow the very weakest examinees (such as the borderline G examinees) to take only the easiest paper, and would therefore provide a more positive experience for them.

The highest boundaries on the easy and main paper in the adjacent levels model were also at a higher proportion of the maximum mark than the highest boundary on the foundation tier paper[9], suggesting that this model could allow more effective use of the mark range for grading.

[9] This contrast would have been more stark if the actual C boundary of 110 had been used.

Of course, the highest boundary on the hard paper was at a lower proportion of the maximum mark than on the higher tier paper (since the easier items had been removed) but it should be borne in mind that only two boundaries were set on this paper, compared to the three boundaries on the easy paper. Inspection of Figure 4 suggests there would be room for an extra boundary on the hard paper to identify the very highest achieving examinees.

Classification accuracy in the different tiering models

Each simulated examinee was assigned various grades as follows:
'True grade' – according to their generating logit ability and the grade boundaries on the bank scale;
'Bank grade' – their grade on a hypothetical test containing all the items in the bank. The raw grade boundaries for this test were derived from its TCC, and the raw scores simply as the sum of the simulated scores on each item in the bank.
'Tiered grade' – according to the simulated score on the foundation tier items for (simulated) foundation tier examinees, and likewise for the higher tier;
'3-level tiered grade' – treating (P1 + P2) as a lower tier and (P2 + P3) as a higher tier and deriving grade boundaries on the aggregate scales corresponding to the bank scale. This is just an alternative tiering model, not the adjacent levels model;
'Adjacent levels grade' – according to the adjacent levels grading rules described above.

Classification accuracy was defined as the proportion of simulated examinees with an observed grade the same as the true grade in the different scenarios, shown in Table 4. The full grade-by-grade cross-classification tables are not shown here to save space.

Table 4: Classification accuracy of different tiering models.

Model	# marks	Classification accuracy[10]
Bank (test comprising all items in the bank)	357	84.2%
Normal tiered model	200	78.1%
3-level tiered model	160	75.6%
Adjacent levels model	N/A (160)	72.0%

Table 4 shows that the main loss in classification accuracy comes from the probabilistic nature of the Rasch model – even a test of 357 marks would 'only' have classified around 84% of the examinees correctly. There was a noticeable drop of 6 percentage points in going from 357 marks to the (actual) 200 marks in the normal tiered model, and a further 2.5 percentage points in going from the actual model to a different tiered model worth 160 marks with three papers targeted at different levels of difficulty. Figure A1 in the appendix shows the test information functions for the bank test, the foundation and higher tiers, and the alternative P1&P2 lower and P2&P3 higher tiers. It is clear that the normal tiering model gave more information than the 3-level tiered model at all grade boundaries, and hence better classification accuracy. Figure A2 in the appendix shows that there was more overlap in the distributions of item difficulty in the P1&P2 lower and P2&P3 combinations than in the normal tiered model so this is not too surprising.[11]

The adjacent levels model gave the lowest value for classification accuracy, a further 6 percentage points below the normal tiered model. This was perhaps to be expected because of the loss of information inherent in basing the grade only on achievement in one of the three papers (i.e. 100 or 60 marks worth) although the presence of the grade B hurdle for achieving the A or A* grade means the total number of marks in the assessment is not so clearly defined.

[10] These statistics are all based on a single simulation. If desired the simulations could be re-run a certain number of times and the distribution of these statistics found.
[11] However, further experimentation showed that even selecting two 160-mark tests based on the easiest and hardest items (not questions) in the bank would not have given more information at the grade boundaries than the normal tiered model. This suggests that without deliberate precise targeting of items at boundaries (which is probably only possible with a large pre-calibrated bank of items), the number of marks (length of test) is the dominant influence on reliability and classification accuracy.

Table 5: Percentage of simulated examinees with a given true grade accurately classified by each model.

Grade	Normal tiered model	Adjacent levels
A*	88.3%	86.7%
A	71.2%	61.6%
B	59.2%	76.8%
C	86.8%	77.5%
D	79.2%	72.2%
E	80.8%	66.2%
F	74.1%	61.3%
G	78.2%	67.4%
U	85.0%	80.8%

Table 5 shows that the adjacent levels model accurately classified a lower percentage of simulated examinees at each grade, except for grade B where the unavailability of this grade on the foundation tier meant that a relatively high proportion of 'true B' examinees obtained a grade C.

Changing the proportion of simulated examinees at each tier

It was noted previously that the cohort on which the simulation was based was potentially unrepresentative of the national cohort in the allocation of examinees to tiers, with around twice as many foundation tier as higher tier examinees. The National Pupil Database (NPD) does not record the tier of entry, but does show the overall distribution of grades across maths syllabuses and exam boards. Table 6 below shows that when the proportion of simulated examinees at each tier was changed such that it was 50-50 (by randomly sampling examinees from the foundation tier to ensure the same number (9,500) as higher tier examinees), the distribution of 'true grade' was closer to the 2011 national distribution.

Table 6: Actual and simulated grade distributions.

Grade	National maths GCSE distribution 2011 (source: Gill, 2012)	This cohort June 2012	'True grades' in original simulated cohort (F=19,000 H=9,500)	'True grades' in modified cohort (F=9,500 H=9,500)
A*	5.4	4.8	5.0	7.1
A	11.5	6.7	6.9	9.3
B	15.6	9.0	13.3	16.7
C	26.9	32.9	22.9	23.9
D	17.0	18.2	19.9	17.2
E	10.8	11.6	17.3	14.1
F	7.4	8.9	8.6	6.8
G	3.9	5.9	4.5	3.6
U	1.6	2.0	1.7	1.3

However, the effect on classification accuracy of modifying the composition of the cohort by tier was small, as shown in Table 7 below.

Table 7: Classification accuracy of different tiering models with different proportions of simulated examinees at each tier.

Model	Original (F:H = 2:1)	Adjusted (F:H = 1:1)
Bank (test comprising all items in the bank)	84.2%	84.3%
Normal tiered model	78.1%	78.7%
3-level tiered model	75.6%	76.3%
Adjacent levels model	72.0%	71.9%

Changing the allocation of simulated examinees to tiers

The simulated examinees were allocated 'perfectly' to tiers by assigning the top half (by generating ability and hence 'true grade') to the higher tier and the bottom half to the foundation tier. The effect on classification accuracy is shown in Tables 8 and 9 below.

Table 8: Classification accuracy of different tiering models.

Model	Original (F:H = 2:1)	Adjusted (F:H = 1:1)	Perfect allocation to tiers (F:H = 1:1)
Bank (test comprising all items in the bank)	84.2%	84.3%	84.3%
Normal tiered model	78.1%	78.7%	81.6%
3-level tiered model	75.6%	76.3%	78.6%
Adjacent levels model	72.0%	71.9%	73.5%

Table 9: Percentage of simulated examinees with a given true grade accurately classified by each model (equal proportions at each tier and perfect allocation to tiers).

Grade	Normal tiered model		Adjacent levels model	
	Original	1:1 + perfect allocation	Original	1:1 + perfect allocation
A*	88.3%	91.8%	86.7%	90.1%
A	71.2%	80.8%	61.6%	69.4%
B	59.2%	80.4%	76.8%	74.2%
C	86.8%	83.6%	77.5%	77.0%
D	79.2%	79.8%	72.2%	71.6%
E	80.8%	79.7%	66.2%	70.0%
F	74.1%	75.4%	61.3%	61.7%
G	78.2%	80.4%	67.4%	67.7%
U	85.0%	85.5%	80.8%	81.1%

The effect of a 'perfect' allocation to tiers was mainly seen in the normal tiered model, where the relatively large number of simulated examinees with a 'true grade' of B were now correctly entered for the higher tier. On the adjacent levels model the largest effect was seen at grade A, where simulated examinees with a 'true grade' of A were now correctly entered for the P2 + P3 combination.

Changing the allocation of items to papers

The method of allocating questions to the three adjacent levels papers was based on empirical information about their relative difficulty, so was in that sense optimal. In practice (assuming no pre-testing) the information about relative difficulty would be based on the judgment of the question setters. To explore the effect of a less-than-optimal allocation of questions to papers in terms of difficulty, the rank-ordering of whole questions by difficulty was changed by adding a random error component and re-sorting. The size of the error component was adjusted by trial and error so that the correlation of the new order with the original order was ≈0.6. This was intended to represent if not quite a 'worst case scenario', at least a reasonably pessimistic estimate (given that this correlation is over the entire ability range). The same whole questions were thus re-allocated to Papers 1, 2 and 3, but the net effect was effectively to swap some easier marks for more difficult ones between Papers 1 and 2, and again between Papers 2 and 3, thus making the papers more similar in difficulty overall, while increasing the spread of item difficulty on each paper.

The re-shuffling resulted in 39 out of 60 marks worth (65%) swapping from P3 (hard) to P2 (intermediate); and 21 out of 60 marks worth (35%) swapping from P1 (easy) to P2. The effect on the grade boundaries and score distributions is shown in Figures 7 and 8.

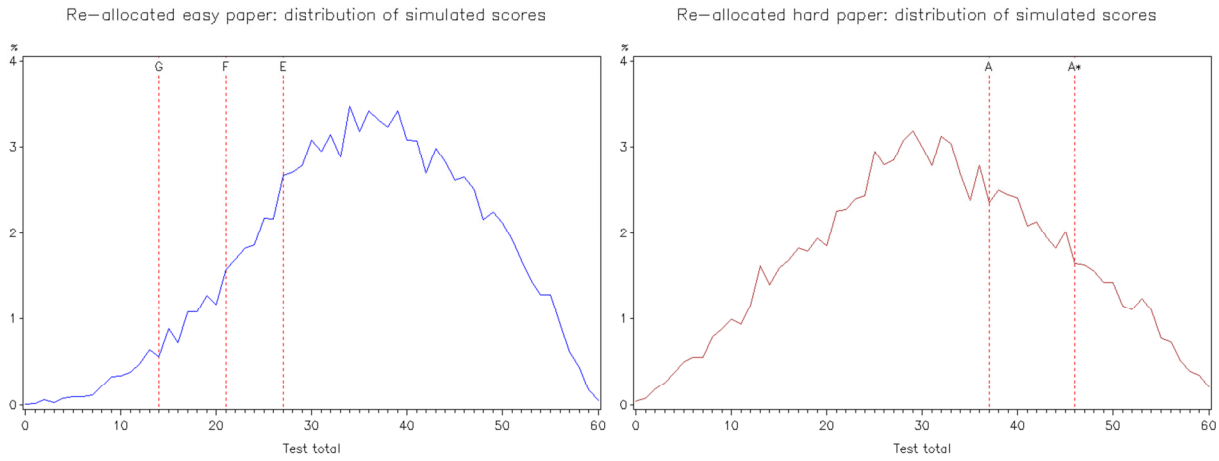

Figure 7: Grade boundaries and score distributions on the simulated 60-mark easy and hard papers after re-allocation of items.

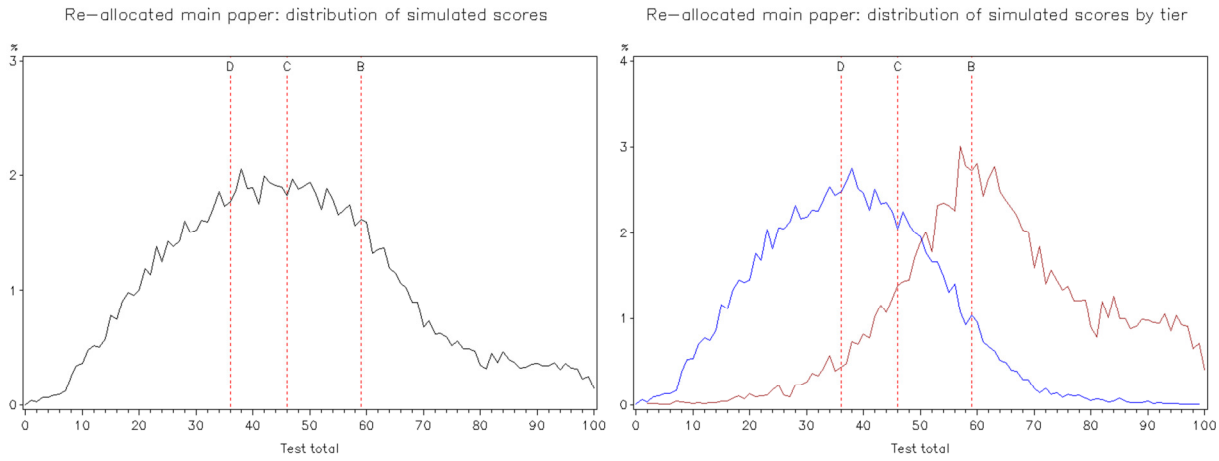

Figure 8: Grade boundaries and score distributions on the simulated 100-mark intermediate paper after re-allocation of items.

Comparing Figures 7 and 8 with the corresponding Figures 4 and 6, it can be seen that the effects of the re-allocation were: i) to move the boundaries closer together on all papers; ii) to move them down on the easy paper (P1); and iii) to move them up on the hard paper (P3). This makes sense, because on all three papers the spread of item difficulties was increased (P1 gained some harder items, P3 gained some easier items, and P2 gained some easier and some harder items). Increasing the spread of item difficulties moves the boundaries closer together. The score distributions on all papers were less skewed.

Table 10: Classification accuracy and consistency after re-allocation of items.

	Accuracy		Consistency
	Original	After re-allocation	Original v after re-allocation
3-level tiered model	76.3%	74.8%	90.0%
Adjacent levels model	71.9%	66.2%	72.1%

The re-allocation of items, representing a mis-targeting of items to tiers, had a small effect on classification accuracy in the tiered model, and a somewhat larger effect in the adjacent levels model. The difference between the two models in classification consistency (examinees obtaining the same grade with both allocations of items to tiers) was much more striking. However, this is not surprising given that the grade in the adjacent levels model is essentially based on

performance in a single paper (with the complication of the P2 grade B hurdle for examinees obtaining A or A* on P3). Therefore the re-allocation of items caused a more substantial change in the effective test taken in the adjacent levels model. On P1, 65% of marks were unchanged, on P2 40% of marks were unchanged, and on P3 only 35% of marks were unchanged. In contrast, on the tiered model for lower tier (P1 & P2) 76% of marks were unchanged and for the higher tier (P2 & P3) 87% of marks were unchanged.

Discussion

Choosing a model for differentiated assessment will require compromises in the extent to which the five desiderata for fitness for purpose identified by Baird et al. (2001) can be met. The results of this study have clearly shown that in terms of psychometric reliability, more (items) is better: both the actual 200-mark 2-tier 2-level assessment and the alternative tiered 3-level 2-tier assessment had better values for classification accuracy than the adjacent levels model. The only apparent weakness in the current model (and this may be a feature of the particular assessment on which the investigation was based) is the relatively large number of examinees entered for the foundation tier who might have achieved a B – in the sense that their foundation tier score was well above a nominal 'B' level on the foundation tier and implied an ability that could have achieved a B (or even an A) on the higher tier. Recent research by Wilson & Dhawan (2014) has suggested that this kind of 'capping of achievement' can be more of a problem on linear assessments like the one considered here than on modular assessments.

However, the conclusion that the simulated examinees in this study had their potential achievement 'capped' by the current tiering model does not immediately follow. First, the analysis assumed that the common grade boundaries had been set in the correct relative place on each tier – i.e. that they corresponded to the same ability. Some early and more recent research (e.g. Good & Cresswell, 1988; Wheadon & Beguin, 2010) has shown that it can be easier to achieve the common grade on the lower tier. Secondly there is the more imponderable question of whether the foundation tier examinees who appeared to have been capable of gaining a B actually would have if they had taken the higher tier paper – for example perhaps they were not taught the higher tier material[12], or were given extensive practice on the easier foundation tier material. This is another way of saying that the assumptions behind the psychometric model might not have held – that the 'missing data' (of foundation tier examinees on higher tier questions) might not be successfully imputed by the model.

In terms of recognising and rewarding positive achievement, the adjacent levels model appeared superior – the lowest grade boundaries on all papers were at a higher proportion of maximum mark, and therefore the question papers gave more evidence about what examinees with a given grade could do (as opposed to could not do). However, this did depend on the difficulty of the questions being well-targeted to the papers. When the questions were re-allocated amongst the three simulated papers to give a less favourable targeting, the grade boundary locations moved closer together, and (on the easiest paper) downwards, reducing some of the potential benefits of this model for inferences about the knowledge and skills of the weaker examinees. This problem (of targeting questions appropriately to grades) was found by Stobart et al. (2005), who concluded that it was a more serious threat to the validity of inferences from the adjacent levels model than the normal tiered model, where less precise targeting is required. Although to some extent true, their discussion seemed to assume that stakeholders would base their inferences on knowledge of the intended grade of the questions. Given that the question papers and mark schemes are published on awarding body websites, arguably it is knowledge of the grade boundary marks and the set of questions that were asked that provides the main indication of what examinees with a given grade know and can do.

[12] In fact, in this syllabus the higher tier examinees would have covered extra (more demanding) content, so (outside of the simulations) the notion of a foundation tier examinee obtaining a grade B does not make sense. This highlights the difficulties in interpreting the meaning of the common grades (C, D and E) obtained on the different tiers in the current tiering model.

As described in the introduction, the 'one route' nature of the adjacent levels model removes the need to achieve comparability of routes, which could be seen as a significant advantage, given the practical and conceptual problems associated with vertical equating. The inclusion of the grade B hurdle meant that the model differed from the Scottish model described in the introduction. The idea behind it would be to ensure that the highest performing examinees – presumably those who might be expected to go on to study maths or related subjects at A level – would have to take two papers (rather than just the highest level paper) and hence that the reliability of their outcome would be higher in that they would have demonstrated grade B achievement on the longest paper and hence would have had their knowledge and skills tested over a wide range of (albeit easier) material. In contrast, the very weakest examinees would only need to enter for the easiest paper if they or their teachers thought there was no chance of them achieving a grade D on the main paper.

However, using the particular adjacent levels scheme simulated here would mean that examinees could achieve a grade B without having to take the hardest paper – which would be inappropriate if those examinees wanted to continue to A level, especially if the syllabus was structured such that there was content on the hardest paper that was not covered on the main paper[13].

Comparing the models in terms of manageability and cost is beyond the scope of this paper, except to note that in principle, if it could be taken first and marked quickly, the scores on the 'main' paper could be used to reduce the amount of assessment required: for example to inform examinees of whether they had a chance of gaining an A or A* (if they had gained a 'B' on the main paper) and thus should take the Paper 3 – or if they needed to take the Paper 1 (if they had failed to get a 'D' on the main paper). Alternatively, if it was deemed more desirable for the weakest examinees to have a more positive experience of the assessment, they could take Paper 1 first, and then make a judgment about whether they should attempt Paper 2 on the basis of how well they did in Paper 1. This kind of immediate feedback for examinee decision-making is not currently possible.

At a more philosophical level, the main difference between the models is in the extent to which one is prepared to accept evidence of achievement outside the expected ability range as validly contributing to the estimate of ability. For example, does exceptionally good performance on easier items give the same information as moderate to good performance on harder items? Using a high-jump analogy, is a jumper who clears a low bar and fails a high bar as good as one who fails the low bar but clears the high one? This question cannot be answered by psychometric modelling alone – the definition of 'a good jumper' has to be fixed in advance. The scoring system for tournament high-jumping presents 'items' of non-decreasing difficulty and the ability estimate is the hardest item 'answered correctly' before three failures. Failures on easier items are largely discounted (three new attempts are available after each success) and only used as tie-breakers to decide a final ranking. If high-jumping used a traditional exam assessment model, each jumper would be given the same fixed set of bars to jump and the winner would be the one who cleared the most. The two models would not necessarily identify the same rank order.

In conclusion, using an adjacent levels model instead of a tiered model would be likely to reduce reliability, but could increase validity – both directly in terms of tying inferences about performance to positive achievement, and indirectly by removing ambiguities about comparability between tiers by having only one route to a given grade. There would be in principle no need for any examinee to have their 'aspirations capped' or to 'fall off the bottom' – with suitable scheduling it could be possible for examinees to enter for all three papers if they were really uncertain about whether they were likely to achieve a grade E or a grade A. The model would be more effective if the three papers could be effectively targeted at different levels of difficulty, such that each paper had a relatively narrow range of item difficulty, but there was a clear progression of difficulty from the easiest to the hardest paper.

[13] Of course, there are various possibilities for how to split the ability range into levels. The choice here was made with an eye to GCSE reform, one of the aims of which is to have more discrimination at the top end. The levels in an adjacent levels tiering model could be structured so that the lowest grade available on the hardest paper was a requisite for progression to A level, and hence an examinee achieving that grade would have covered all the necessary content.

References

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

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521-532.

Gill, T. (2012). GCSE uptake and results, by gender 2002-2011. Cambridge Assessment Statistics Report Series No. 49. http://www.cambridgeassessment.org.uk/ca/digitalAssets/205874_Report_49_-_GCSE_uptake_and_results__by_gender.pdf Accessed 15/04/13.

Good, F. J., & Cresswell, M. J. (1988). Grade awarding judgments in differentiated examinations. *British Educational Research Journal, 14*(3), 263-281.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Stobart, G., Bibby, T. & Goldstein, H. (2005). *Moving to two-tier GCSE mathematics examinations. An independent evaluation of the 2005 GCSE Pilot and Trial.* London: Institute of Education.

Wheadon, C., & Beguin, 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.

Wilson, F. (2013). *Exploring a 'one route' adjacent levels model: GCSE mathematics (1969) pilot.* Cambridge Assessment internal report.

Wilson, F. & Dhawan, V. (2014). *Capping of achievement at GCSE through tiering.* Paper presented at the British Educational Research Association conference, London, September 2014.

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

Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (4 ed., pp. 111-153): ACE/Praeger series on higher education.

Appendix

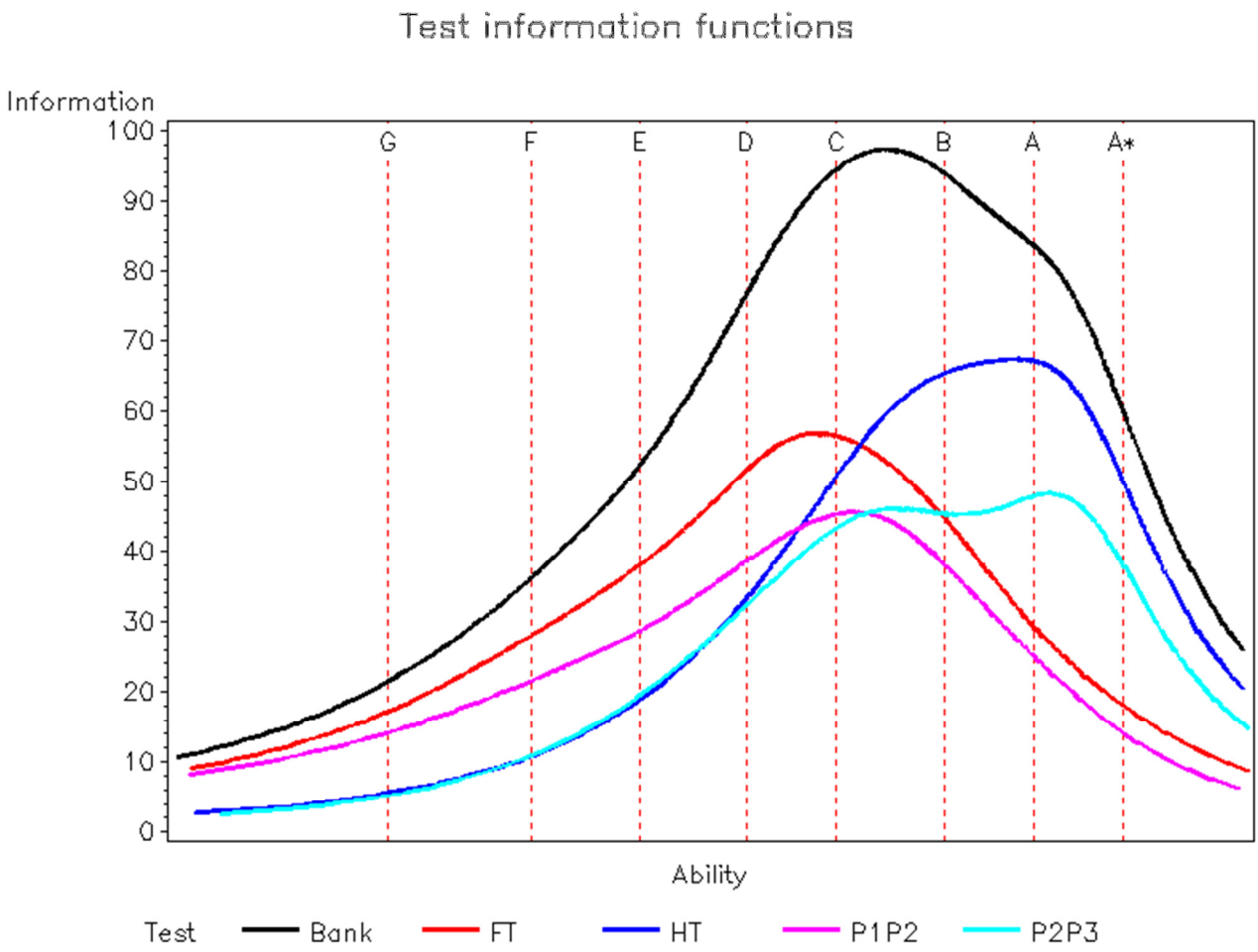

Figure A1: Test information functions for the 357-mark bank test, the 200-mark foundation and higher tiers, and the 160-mark P1&P2 and P2&P3 combinations.

The test information at a given ability is the sum of the item information functions at that ability across all items in the test (see for example Yen & Fitzpatrick, 2006). It is equal to the inverse of the square of the standard error of the ability estimate.

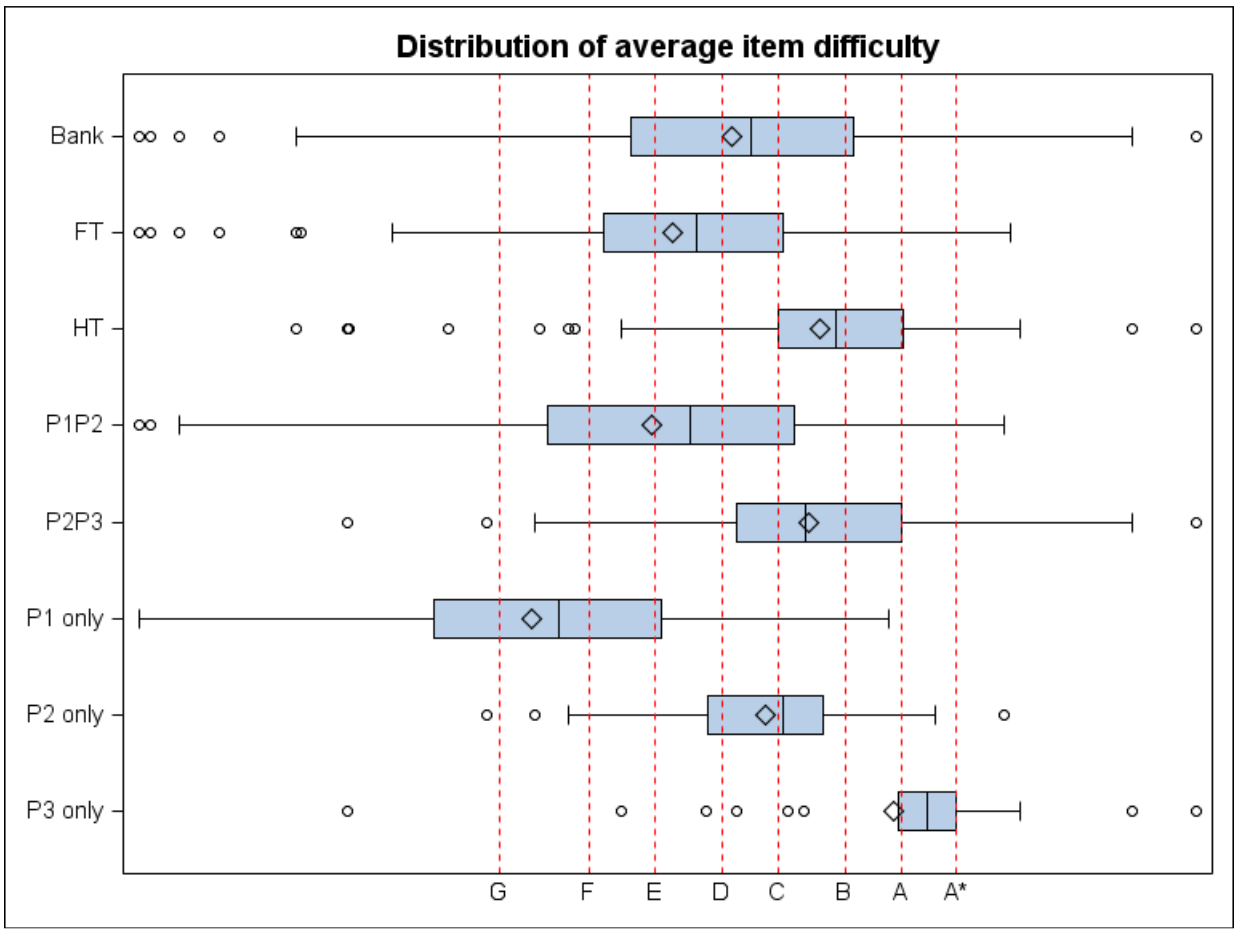

Figure A2: Distribution of item difficulties for the 357-mark bank test, the 200-mark foundation and higher tiers (FT & HT), the 160-mark P1&P2 and P2&P3 combinations, and the individual P1, P2 and P3 papers.