Technical guidance for schools on estimating exam grades, summer 2020

By ASCL member and Executive Headteacher David Blow, May 2020

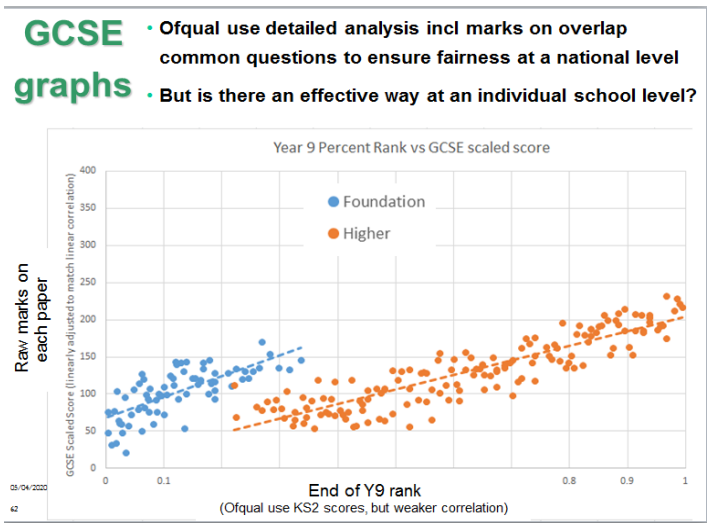
This paper follows on from Duncan Baldwin's ASCL paper *Coronavirus: Emerging principles and guidance regarding teacher-assessed grades for summer 2020* (from Monday 30 March), and the most recent guidance *Coronavirus: Guidance regarding centre-assessed grades for summer 2020* (Thursday 9 April).

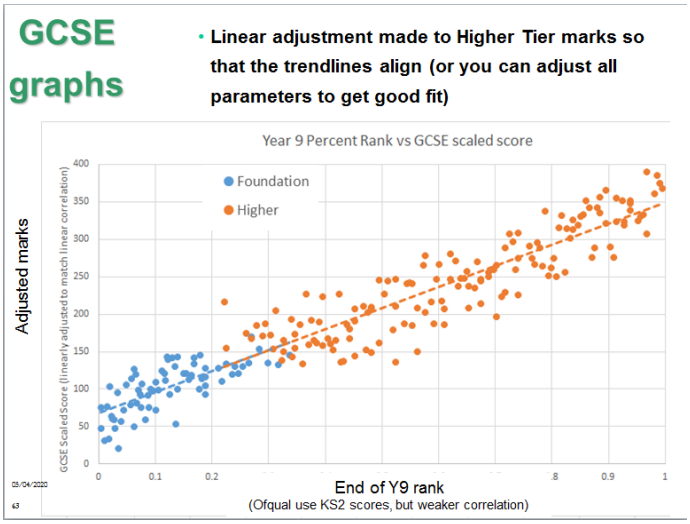
The details for the techniques including mathematical and spreadsheet formulae are in the Annex at the end of this paper. There are also three example spreadsheets to go with this to help illustrate the approaches.

1. Technique for combining different papers within test to give single overall – example Maths. Needed for tiering, but also useful in more general situation where different papers have been set covering a cohort.
2. Subjects with a practical component, where they may be substantial difference in candidate performance in written and practical papers – example: music
3. Techniques for comparing performance across different groups, including setting – example: English

1. Techniques for combining different papers within test to give single overall mark

An example spreadsheet called GCSE Maths tiered example.xlsx is provided.





a) The left-hand graph above shows the results of the two tests (Foundation and Higher) ("raw marks") plotted against an overall ranking measure for the cohort, say their performance in a range of subjects at the end of KS3, or even KS2 / CAT scores.
b) You can see how in the graph on the left, quite rightly, the Foundation marks cover a broad spread, as do the Higher ones, and for each tier you can calculate a trendline / line of best fit, but that as it stands with raw scores, pupils in the overlap are getting a higher raw score in Foundation than Higher
c) Using the techniques explained in the Annex and in the example spreadsheet, the raw scores of the Higher paper are mathematically adjusted by a formula, so that they increase in such a way that the adjusted trendline for Higher is an extension of the trendline for Foundation. That way, the overlap students are getting similar adjusted marks whether they sat the Foundation paper or the Higher paper. There is a now an overall mark across the year group.

d) This is displayed in the right-hand graph above, where the y-axis is "adjusted marks"

Details of the techniques are in the Annex and in the example spreadsheet "GCSE Maths tiered example.xlsx"

2. Subjects with a practical component, where they may be substantial difference in candidate performance between written and practical papers – example: music

An example spreadsheet called GCSE Music incl practical.xlsx is provided.

For subjects with just written papers, there is usually a good correlation between the scores in each paper, and so the marks and evidence assembled are likely to cross-support, giving more confidence to the final estimated mark, even if there are some questions over particular parts of the evidence.

However, this is often not the case for practical subjects, where there can be a number of students who are good at practical but weaker at written, or good at written but weaker at practical.

This means that it can be more difficult to get an accurate estimate of their likely grade, as it may be distorted by incomplete evidence unless appropriate steps are taken to adjust correctly.

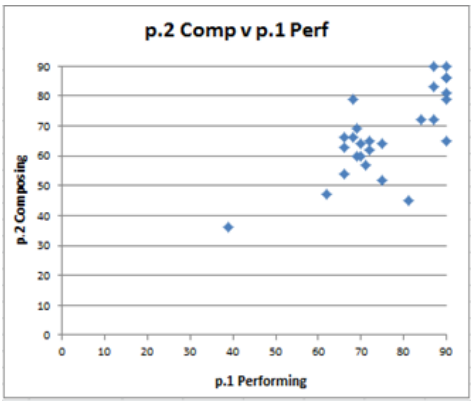
Step 1
Note the percentage weighting for each of the components for the final exam.

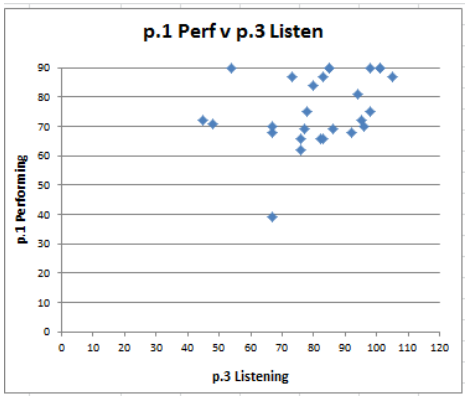
Step 2
Examine the evidence you have available for each component for this year. How robust is it? If you have sufficient well-balanced evidence for each component, then you are in a good position, and may not need to do the following steps.

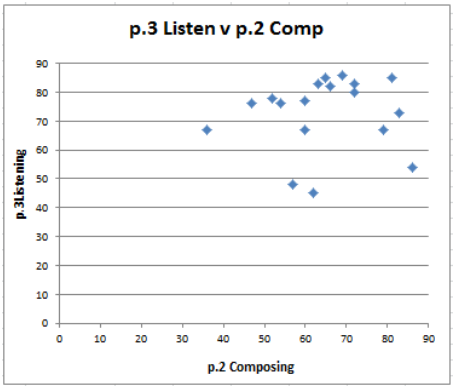
Step 3
Look back at previous years' component marks and analyse them as described in the Annex. This will identify those components whose marks tend to be in line with each other (i.e. there is a reasonable correlation between those components), and those where the marks can be quite different for individual pupils (i.e. there is low correlation between those components).

For example in GCSE Music there are 3 papers: Paper 1 - Performing, Paper 2 - Composing; Paper 3 Listening, with weightings on 30% : 30% : 40%. You can see that there is a broad link between p.2 Composition and p.1 Performance (l-h graph below), but no correlation between p.3 Listening and either of the other papers (centre and r-h graphs), yet it carries 40% of the marks.

So, whereas the marks in p.2 Composition and p.1 Performance will support each other, we need to make sure that we have collected as much evidence as we can for p.3 Listening to make sure that the performance of the candidates in that paper is reflected in the estimated grades. This may well mean using mixed or incomplete data, but that is where it is important to examine the different options and their outcomes and then to form a reasoned judgement about how the estimate is to be calculated. See the example spreadsheet for details.







Details of the techniques are in the Annex and in the example spreadsheet ""GCSE Music incl practical.xlsx"

3. Techniques for comparing performance across different groups, including setting – example: English

English (as a large setted subject without tiering)

A sample set of data from English GCSE English setted marks.xlsx is supplied with fve sets of marks and an average grade from KS3 where 1 is high.

There a parallel band structure in this school (X and Y) and English is taught in each band in 2 parallel top sets, set 2 and set 3.

Questions in looking at the data

Q.1 Look at the numbers taking each test / assessment, the average mark and standard deviation (or spread), and the link (correlation) with the KS3

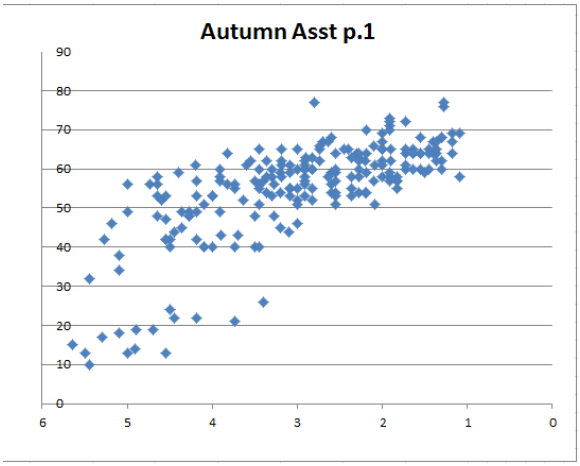
- The class test (col D) has the strongest correlation with KS2 and from the standard deviation we see that it has a good spread of marks. However, unfortunately, the marks for one set are missing, so will probably have to discard this data, and there are still 4 sets left

GCSE English setted marks.xlsx

	A	B	C	D	E	F	G	H	I
1		Pupil numn ber	set	Class test	Autumn Asst p.1	Autumn Asst p.2	Jan Mock p.1	Jan Mock p.2	*KS3 ave gde (1 high)*
2									
3	number	200	200	165	199	198	200	199	*199*
4	ave			63.6	55.3	59.0	76.4	60.7	*2.9*
5	std dev			15.9	13.1	8.4	14.2	10.4	*1.1*
6	correlation with KS3 ave gde			0.67	0.53	0.58	0.60	0.46	*1.00*

Q.2 Plot graphs of each of the tests, with the marks on the y-axis and the prior attainment across the x-axis

- Looking at the graph for Autumn Assessment p.1, it is clear that something went awry for X set 3, either in terms of the marking or the taking of the test. It might be possible to make an adjustment to their marks to bring them up to the level of Y set 3. When doing graphs, it can be very useful to plot each class as a different series with a different colour



Q.3 Construct grids / tables looking at the mark distribution against class for the total and for each tests.

- The table on the right shows you the distribution of marks for the January Mock p.2. It would appear that there is inconsistent marking between the teacher of set X1a and the others. The marking of the set 2s and set 3s looks reasonably consistent. There probably should be an adjustment to the marks of set X1a for this test.

Jan Mock 2		number of pupils in mark range					
min	0	50	60	70	80	90	
max	50	60	70	80	90	100	
X/En1a	0	2	5	17	5	2	31
X/En1b	0	0	1	3	18	9	31
X/En2	2	6	11	12	1	0	32
X/En3	8	5	1	1	0	0	15
Y/En1a	0	0	1	11	12	8	32
Y/En1b	0	0	0	5	19	9	33
Y/En2	1	7	8	10	2	0	28
Y/En3	7	3	3	1	0	0	14
	18	23	30	60	57	28	216

The example spreadsheet "GCSE English setted marks.xlsx" gives examples and formulae in order to carry out these simple but effective tests

These examples illustrate the value of doing simple analysis whether through graphs or scattergraphs on the consistency and quality of the data.

Annex with detailed technical information

1. Techniques for combining different papers within test to give single overall mark (Maths)

An example spreadsheet called "GCSE Maths tiered example" is provided.
You can see from the screenshot on the right that the candidates are listed in order, firstly by Tier, and then a Y9 rank figure in column E. This is needed to give a reference scale, but its precise nature is not critical. This uses the average across subjects at the end of Year 9, and then given a ranking score on a scale from 0 to 100.

GCSE Maths tiered example v1.xlsx

	B	C	D	E	F
1	Pupil no	Tier	Raw Mark on paper	Y9 rank	Adjusted mark
2	1	Foundation	48	0.4	48
3	2	Foundation	76	0.4	76
4	3	Foundation	32	0.9	32
5	4	Foundation	77	1.4	77
6	5	Foundation	34	1.7	34
7	6	Foundation	104	1.9	104
8	7	Foundation	64	2.1	64
9	8	Foundation	61	2.4	61
10	9	Foundation	60	2.6	60
11	10	Foundation	48	2.8	48
12	11	Foundation	96	3.3	96
66	65	Foundation	135	26.4	135
67	66	Foundation	171	26.9	171
68	67	Foundation	154	28.3	154
69	68	Foundation	135	29.8	135
70	69	Foundation	133	31.7	133
71	70	Foundation	146	33.6	146
72	71	Higher	112	22.1	218.0486
73	72	Higher	69	22.5	155.991
74	73	Higher	83	25.9	176.1958
75	74	Higher	78	26.9	168.9798
76	75	Higher	90	27.8	186.2982
212	211	Higher	191	95.6	332.0614
213	212	Higher	193	96.1	334.9478
214	213	Higher	175	96.6	308.9702
215	214	Higher	232	96.6	391.2326
216	215	Higher	212	98	362.3686
217	216	Higher	229	98.5	386.903
218	217	Higher	222	99	376.8006
219	218	Higher	217	99.5	369.5846

For the two sets of marks (Foundation and Higher), you can calculate and plot a line of best bit / trendline using the following formulae to calculate the slope and intercept on the y-axis of the line s:
=SLOPE(D2:D71,E2:E71) =INTERCEPT(D2:D71,E2:E71)
with the values being given in the table on the right.

H	I	J	K	L	M
		Found-ation	Higher		
Slope of best fit (m)		2.828	1.960	ratio =	1.443
Intercept of best fit (c)		67.74	7.85		

We now want to adjust the marks so that all the marks (Foundation and Higher) have a single trendline. The line of best fit for the Higher paper is $y = 1.96 x + 7.85 = = m_{high} x + c_{high}$
so we use

$y_{adjusted} = m_{adj} y + c_{adj}$ where $m_{adj} = m_{found} / m_{high}$ and $c_{adj} = c_{found} - (m_{adj} \times c_{high})$

because

$y_{adjusted} = m_{found} / m_{high} (m_{high} x + c_{high}) + c_{found} - (m_{adj} \times c_{high})$

$= m_{found} x + c_{found}$

i.e. is the same line as the foundation one

So, the formula for the "Adjusted Mark" in cell f75 for example multiplies the raw mark by 1.443 and adds 56.4 to make it equivalent to a mark on the Foundation paper.

O	P	Q	R
Linear Scaling for Raw to Adjust. Mark			
		slope	const.
	Foundation	1	0
	Higher	1.443	56.4

GCSE Maths tiered example v1.xlsx

	A	B	C	D	E	F
1		Pupil no	Tier	Raw Mark on paper	Y9 rank	Adjusted mark
2		1	Foundation	48	0.4	48
3		2	Foundation	76	0.4	76
4		3	Foundation	32	0.9	32
5		4	Foundation	77	1.4	77
75		74	Higher	78	26.9	168.9798
76		75	Higher	90	27.8	186.2982

These figures are then used to plot the graphs. A subtle point is that detailed examination of mock papers may show that a small disconnect at the overlap point may better match against grade boundaries, but this method allows for that fine tiering based on the knowledge of the teacher.

2. Subjects with a practical component, where they may be substantial difference in candidate performance between written and practical papers – example Music

Music (as a practical subject)

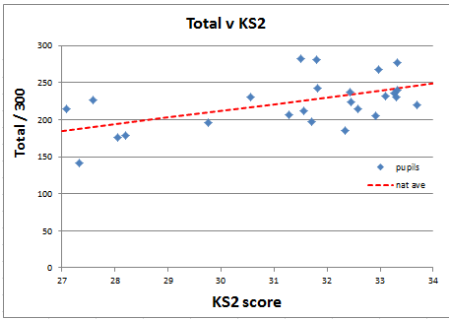
Whereas a subject like Maths would usually see a strong correlation between two papers, the situation can be quite different in a practical subject especially where there is some kind of "performance" element. An example spreadsheet is provided for Music to illustrate some of the issues. Although details will vary between boards and specifications, in general, there are three papers: Paper 1 - Performing, Paper 2 - Composing; P.3 Listening, with weightings on 30% : 30% : 40%. The first two papers are done as Non-Exam Assessment (NEA), and were likely to be in various stages of completion varying from pupil to pupil and school to school.

What evidence is likely to be available?

P.3 Listening (40%) is done as a formal exam and so there is likely to be a Y11 mock exam in the last few months as well as other exams from earlier in the GCSE course, as well as practice tests in class.

For Paper 1 - Performing and Paper 2 - Composing (combined 60%), the evidence base will be much more variable even within a school with pupils at different stages of completing and handing in the latter and doing the performances for the former. There may be some mocks or practice pieces from earlier in the course.

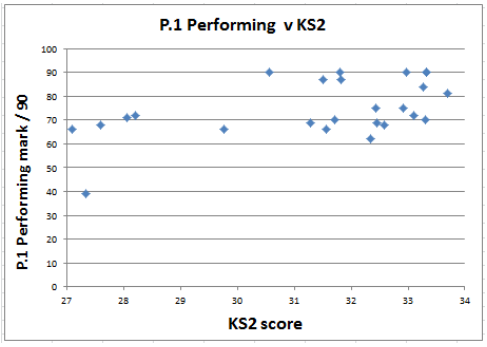
What can looking at previous years' data show?

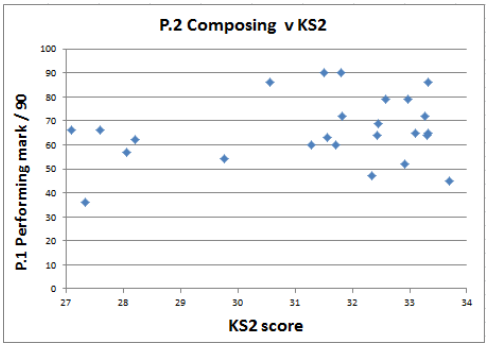


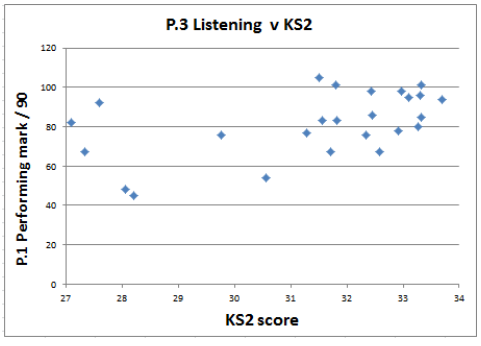
The spreadsheet is a complete set of data for 27 pupils from the actual GCSE results. The KS2 score is given and is used in the same way as the Y9 rank figure was used in the example above. Also, though we know that the KS2 prior attainment of the group will be used to check the grades being allocated to the group, so this gives an additional feature to look at, and see what the issues might be. But this example illustrates the importance of getting a picture (ASCL Leadership of Data Conference delegates will be familiar with the Mona Lisa analogy of comparing the beauty of the whole picture with the uniform dark green rectangle on its chromatic average - message: look at the picture not just the average).

Even a quick glance at the scattergraph shows that there is quite a spread at an **individual** level. And it's important to stress that the Transition Matrices analysis will show that in overall terms, there's a good fit for the **group** between the grades which were actually awarded and those which would arise from the TM model. The challenge is to make sure the right pupils get the right grades.

You can see from the graphs below for each component against KS2, that there is even lower linkage for individuals for each component. Listening, which is the exam paper, is the one which shows a little, probably because it is more akin to skills in English etc and can be revised for etc.



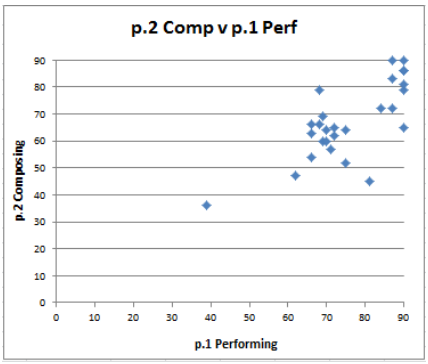


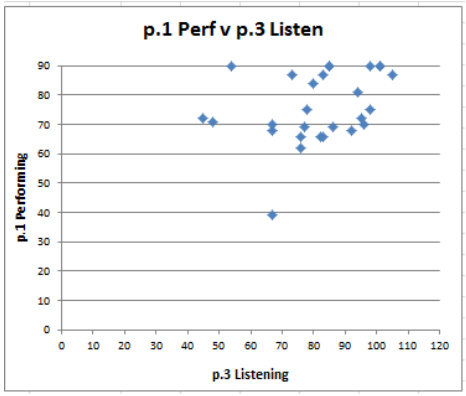


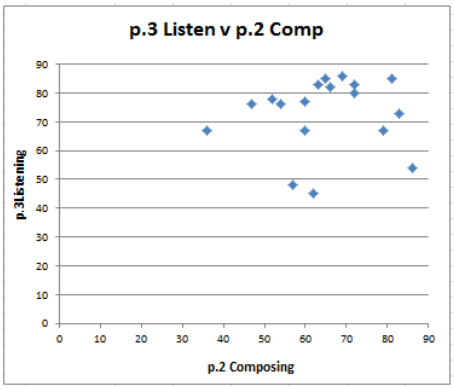
This is quantified in the table on the right of the correlation (r-squared) between the different variables. The figure in the table represents how well can you predict the outcome knowing the input.

Correlations (r-squared)					
	p.1 Performing	p.2 Composing	p.3 Listening	Total	KS2 score
p.1 Performing	1.00	0.56	0.11	0.70	0.04
p.2 Composing	0.56	1.00	0.09	0.70	0.00
p.3 Listening	0.11	0.09	1.00	0.52	0.17
Total	0.70	0.70	0.52	1.00	0.07
KS2 score	0.04	0.00	0.17	0.07	1.00

For example, looking at the graph of Paper 1 - Performing against Paper 2 - Composing, you can see that there is abroad linkage - high performance in one links to high in the other, middle to middle and low to low, so the figure of

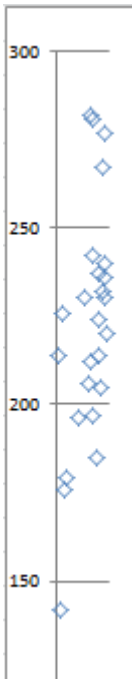






0.56 means 56% of one variable can be estimated from the other. Each paper will have a link with the total score because each forms a roughly equal part (30:30:40) of the whole.

But the fact that there is almost no correlation between p.3 Listening and either of the other two p.1 or p.2 means that we need to get good evidence for p.3 so that it can make its contribution to the whole. Whereas the link between p.1 and p.2 mean that evidence for the two papers can be brought together to support each other.



And fortunately, in the case of music, this may well be the case as described in the section above "What evidence is likely to be available?"

This paper is focussing on the issues regarding marks, but just to look ahead for a moment, squashing the x-axis as in the graph on the right highlights the spatial distribution of the marks, showing the clustering and spacing, which will be critical when deciding which grades to allocate to whom. And we will show in that next Guidance how the Transition Matrices approach can be used effectively to get a robust and fair overall distribution, whilst allowing the fine-tuning to give the right grades to the right pupils.