Improving Selection to the Foundation Programme

Appendix H

Report of the EPM pilot

Educational Performance Measurement (EPM)
Report of 2010 Pilot of the EPM draft framework

This document summarises what has been done to pilot the production of EPM scores by UK medical schools, the findings from the pilot, and suggested next steps. The current version is for consideration by the Improving Selection to the Foundation (ISFP) Project Group.

Table of Contents

1. Background

Applicants to the Foundation Programme receive a score based on performance at medical school in relation to their cohort, as ranked into four quartiles worth 40, 38, 36 and 34 points. This score is then combined with their score from an online application form, white space questions (total of 60 points), including 10 points for degrees, prizes, publications and presentations.

Concerns about the use of academic quartiles, raised before and during an extensive and detailed Option Appraisal[1], are based around the comparability between applicants from different medical schools and discrimination of applicants at the margins between quartiles. Given issues around the lack of transparency and lack of consistency across medical schools, one of the main drivers to review the quartile system is to ensure defensibility in the event of legal challenge from an applicant. Stakeholder feedback showed strong support for the use of some measure of academic performance as well as non-academic and possibly extra-curricular activities. The advisory international Expert Panel supported the principle of making greater use of information accumulated during medical school, and the development of a standardised measure of educational performance.

The recommendations of the Improving Selection to the Foundation Programme Steering Group were to pilot and evaluate:

1. A Situational Judgement Test (SJT) to assess professional attributes, judgement and employability

In combination with

2. An Educational Performance Measurement (EPM) of applicants at their medical school to assess clinical knowledge/ skills as well as wider personal achievement

The EPM as a selection tool refers to a differential ranking score produced by the applicant's medical school to reflect the applicant's achievements or performance on a range of assessments in relation to their cohort up to the point of application to the Foundation Programme. It was envisaged that the EPM would be derived using a specified and standardised transparent framework of existing performance measures, with a weighting agreed with medical schools in consultation with students and other stakeholders. All UK and non-UK medical schools would be required to provide a local educational performance ranking to the UKFPO derived using the standardised framework.

The ISFP Steering Group envisaged that a uniform transparent framework for an EPM would address some of the current concerns about comparability between applicants of the same quartile from different schools, and that it would enable greater granularity[2]. Depending on the results of the 2010-11 pilots, it is anticipated that the EPM would be used in combination with a second selection tool, resulting in a combined higher level of granularity than would need to be achieved by the EPM alone.

2. Draft quality criteria

The following quality criteria are measures, proposed by the ISFP project team for consideration by the Project Group, against which to evaluate the Educational Performance Measurement (EPM):

- The EPM is a reliable and representative measure of the applicant's performance in their educational progression at medical school up to the point of application
- The EPM is a valid measure in relation to selection to the Foundation Programme (i.e. it is measuring factors that have a bearing on suitability for the job)
- The EPM is a sufficiently granular measure
- The EPM is a fair measure

[1] Medical Schools Council (2009) *Selection to Foundation: An Option Appraisal*

[2] Granularity is taken to mean the differentiation between applicants (i.e. the number of different scores achieved)

- The EPM is not overly expensive to administer and quality assure
- Medical schools are able to adhere to the framework, and introduce it within suitable timescales
- Where medical schools are not able to produce a reliable EPM, there is a legal and justifiable fall-back
- There is no requirement on any medical school (UK or non-UK) to modify its curriculum in order to comply with the framework
- The framework complies with all relevant legislation

3. Methodology

Developing the pilot EPM framework

Two in-depth consultations in autumn and winter 2009 involving all 30 UK medical schools whose students apply to the Foundation Programme[3] informed the development of the draft EPM framework for piloting.

The first consultation, in September and October 2009, gathered evidence around the information currently used to inform quartile rankings, and the assessment information on student performance currently collected and utilised by UK medical schools. The range of information – and the formats – varied widely between schools, as follows;

- The survey revealed a range of assessment types, including clinical skills, curriculum knowledge, Student Selected Components and measures of professionalism, as well as summative, formative and progress-testing. The timing of some assessments is variable between schools.
- The two areas where all schools collect information on student performance are clinical skills and curriculum knowledge, although the number of assessments of each – and the relative weighting (e.g. within course compared with end of stage) – is variable, both within and between schools.
- Between 7% (Manchester) and 100% (Cambridge, Imperial, Oxford) of students at a given school intercalate (n.b. graduate entry students already hold a degree), with a median of 25%. Intercalation is usually competitive entry, based on performance at medical school, and not available to all students (for example graduate entrants).
- The granularity of information available is variable between schools and between assessments within schools. A full ranking within a cohort is sometimes neither feasible nor desirable, as assessments are designed to assess competence, not a spread of marks. Furthermore, as a matter of university policy many schools retain only the grades and bandings awarded for performance on assessments, rather than retaining the more-detailed raw scores. It is difficult to quantify scores for Student Selected Components (usually pass/ fail), measures of professionalism (can be pass/ fail) and extra-curricular activities (rarely collected systematically by the school, impossible to quantify).

The second consultation, in November and December 2009, consulted around the principles for an EPM framework, the weighting that might be awarded to curriculum knowledge and clinical skills, the proposal for two sets of quartile scores (one for written assessments, and the other for practical assessments), and how additional points for prizes, publications, presentations and previous degrees might be awarded. The survey responses are summarised below:

- All responses recommended that schools should determine how many assessments to include, and the weighting of these, with a general consensus that the weighting between clinical skills and curriculum knowledge should be 50:50.
- Most responses agreed that there should be a spread of scores across all years, with a higher weighting in the later/clinical years. Differences between Graduate Entry (GE) (4 years) and standard entry courses (5/ 6 years) should be taken into account.
- Responses indicated the need to take into account performance in both written tests of knowledge and those of clinical skills exams. However, some concerns were raised about separating scores for these two areas.

[3] Excludingthe University of St Andrews. Students transfer to the University of Manchester or to other Scottish medical schools at the end of Year 3, ahead of their application to the Foundation Programme.

- More than four fifths of schools believed that the EPM should award points for presentations, prizes and publications, in addition to points for performance in relation to the cohort. School responses also suggested that additional points could be awarded to recognise the top performing students within the cohort. Responses emphasised that guidance regarding the award of additional points should be standardised and transparent.
- 13 schools (7 GE) thought that there should be additional points for previous degrees as well as those gained during medical school; 7 (4 GE) thought that points should only be awarded for degrees gained during the time at medical school; 6 (3 GE) thought that no points should be awarded for degrees.

A working draft EPM framework for piloting, informed by the feedback from the two consultations, was agreed by the ISFP Project Group in May 2010 and by the Medical Schools Council in June 2010. The draft EPM framework called for students to be given two scores - one relating to written assessments (as an approximation for curriculum knowledge), and the other relating to practical assessments (as an approximation for clinical skills) – according to a specific prescription of weightings between earlier and later years of the course. The two scores would then be combined to provide an overall score.

The pilot EPM framework is attached as Appendix A.

Pilot EPM data collection

UK medical schools were asked to produce a pilot EPM, based on clinical skills and curriculum knowledge up to the point of application to the Foundation Programme 2010 (FP2010), for each of their students who had applied to the FP2010. As the pilot used retrospective data, the EPM scores could then be analysed with the original quartile scores supplied by the medical school, and with the application form scores achieved by applicants during the last recruitment round. At this stage, the framework did not include other aspects of performance such as Student Selected Components (SSCs), academic excellence, previous degrees and/or prizes, publications and presentations, as there was either no consensus around scoring (SSCs, academic excellence) or the information already existed for the cohort (previous degrees, prizes, publications and presentations).

Named contacts at each medical school were sent an individualised Excel template for completion, with the names and email addresses (unique identifier) pre-entered. Schools were tasked to complete the data manipulation, according to the prescription, within one calendar month and to return the data to Jobsite[4]. A two week extension was granted to some schools, with the final anonymised dataset provided to the Medical Schools Council in mid-August.

EPM data were provided as normalised scores, such that comparisons could be made within a single school cohort. The following analysis is based on the EPM scores provided for 5373 applicants to the FP2010, from 22 UK medical schools. Three schools provided EPM scores for a further 442 students in different cohorts, although the scores could not be correlated with other information. Analysis of the EPM data includes the cohort which graduated in 2010, and a small number of applicants who had deferred their application for 1 or 2 years. This analysis excludes applicants to the academic Foundation Programme who did not complete the white space question application form, as there is a separate application process.

The scores were then converted into two sets of quartile ranks reflecting performance in i) written assessments of curriculum knowledge (written EPM scores) and ii) practical assessments of clinical knowledge (practical EPM scores).

Schools were also invited to report back on the issues they encountered in following the pilot EPM framework. A summary of this feedback is attached as Appendix B.

[4] Jobsite holds application data for the Foundation Programme, by contract with the UKFPO. Anonymised data were correlated and provided to the Medical Schools Council free of charge for the purpose of this pilot.

4. Findings of the EPM pilot

EPM scores were produced by 25 of the 30 UK medical schools with graduating final year students within the timeframe of the pilot. 22 medical schools provided data in a format that enabled analysis of EPM scores with existing data on quartiles and scores on the white space questions. Divergences from the prescribed EPM framework are detailed in Appendix C, and should be referred to when interpreting the EPM data.

Schools reported that the framework required the use of additional assessment data (for example early years assessments), whereas other schools reported that the framework limited the number of assessments that could be used (for example by requiring a split between 'written' and 'practical' EPM scores). EPM scores were provided in raw score format, and Figures 1 and 2 illustrate the range of marks and formats of scores. Whereas in Southampton, written EPM scores consisted of 13 possible scores, for 228 applicants, between 0.5 and 2 (2 being the lowest performing applicant and 0.5 the best performing applicant), Edinburgh provided written EPM scores enabling a complete ranking of all 208 applicants between 65.8 and 88.8 (to 6 decimal places), and Sheffield written EPM scores were provided as 210 possible scores (223 applicants) within a range of 139.6 and 397.6.

8 of 25 schools provided data on written EPM scores which enabled a full rank (no tied scores). In other schools as many as 100 of 228 students had a tied score. For 12 of 25 medical schools, applicants could be ranked into equal quartiles using written EPM scores. Owing to tied scores, for the remaining quartile groups – with the exception of Southampton – there was a variation of between 4 and 10 individuals in the group size.

Figure 1: Mean, median and mode of written EPM scores (25 schools)

School	No. of students	Lower range	Upper range	Mean (2.d.p)	Median	Mode	Mode (no of ties)
Aberdeen	173	9.78	18.33	13.32	13.87	11.67	3
Barts and The London	322	50.28	86.33	66.11	66.71	No ties	No ties
Birmingham	382	51.57	80.00	64.95	64.40	No ties	No ties
Brighton and Sussex	125	57.90	87.92	71.11	71.53	No ties	No ties
Bristol	243	53.02	76.67	63.58	63.31	No ties	No ties
Cambridge	135	35	60	48.90	49	51; 52	12
Cardiff	314	37	86	55.95	77	71.35	9
Dundee	150	31.8	43.9	36.55	36.35	33.60; 34.90; 39.70	3
Edinburgh	208	65.82	88.84	77.22	77.09	No ties	No ties
Glasgow	220	17	42	13.88	28	29	27
HYMS	116	195	279	236.39	239	242; 247	6
Keele	105	602	1024	690.62	907	842; 850	4
King's College London	374	52.95	81.16	68.00	67.71	No ties	No ties
Leeds	261	50.57	94.50	71.77	71.50	66	9
Leicester	203	44.6	86.9	63.33	62.6	60	4
Liverpool	312	66.00	83.03	73.36	74.82	76.1	5
Newcastle	350	12.5	28	22.73	23	21	74
Nottingham	313	51.19	82.17	66.03	65.89	No ties	No ties
Oxford	158	180.83	217.16	199.47	199.70	No ties	No ties
Sheffield	223	139.60	397.60	303.64	302.60	273.5	3
Southampton	228	0.50	2.00	1.26	1.30	1.3	100
St Georges, London	294	37.33	83.82	67.10	66.80	60	4
UCL	329	42.74	89.06	68.21	69.07	69.08	2
Uni. of East Anglia	118	148.60	255.40	210.00	213.70	224.4	3
Warwick	159	1.67	43.75	27.54	29.17	30	15

Three schools provided data which enabled a full rank (no tied scores) for practical EPM scores. For 16 of 25 schools, applicants could be ranked into four equal quartiles using practical EPM scores. Owing to the number of tied marks for the remaining quartile groups there was a significant variation, for example in Newcastle, four times as many applicants were placed in the third quartile (126) as in the fourth (32). For two schools (Leeds and Southampton), there was insufficient granularity based on the practical EPM scores to create quartile bandings.

Figure 2: Mean, median and mode of practical EPM scores (25 schools)

School	No. of students	Lower range	Upper range	Mean (2.d.p)	Median	Mode	Mode (no of ties)
Aberdeen	173	9	20	13.68	13.86	14	8
Barts and The London	322	50.00	75.25	62.00	62.49	No ties	No ties
Birmingham	382	45.53	78.41	63.39	63.52	67.54; 68.41	2
Brighton and Sussex	125	54.00	83.40	70.57	71.50	62	14
Bristol	243	53.63	74.21	63.67	63.67	59.56; 61.61; 62.19; 62.70; 66.16	2
Cambridge	135	108	161	138.20	137	140	10
Cardiff	314	35	82	67.60	68	70	22
Dundee	150	65.30	85.40	73.12	73.10	66.3	3
Edinburgh	208	63.54	90.90	79.53	79.89	No ties	No ties
Glasgow	221	10	28	22.25	22	24	46
HYMS	116	46	75	58.94	58	56	12
Keele	105	539	796	658.70	660	626; 660; 663; 666; 727	4
King's College London	374	50.39	84.83	66.67	67.31	68	3
Leeds	261	0	100	64.52	60	60	131
Leicester	203	31.20	85.40	69.42	69.80	69	6
Liverpool	312	66.84	82.55	72.94	72.80	67	5
Newcastle	350	19	28	23.67	24	24	81
Nottingham	313	48.28	82.87	65.61	65.25	67.52; 62.32	3
Oxford	158	71.28	89.67	80.06	79.97	No ties	No ties
Sheffield	223	350	572	483.40	488	483	6
Southampton	228	0.5	2	1.30	1.5	1.5	115
St Georges, London	294	56.66	88.48	71.04	71.36	60	6
UCLMS	329	53.73	89.51	74.21	76.56	79.37	2
Uni. of East Anglia	118	712	914	821.40	825.25	817.5	3
Warwick	159	20.83	50.00	43.82	45.83	50	15

The number of assessments used to derive a measure of educational performance, the range of possible scores for each assessment, and the weightings applied to different assessments, inevitably affects the range of total scores achieved by individual applicants. For example, the University of Oxford included 14 assessments (8 written, 6 practical), whereas the University of Southampton included 3 assessments (2 written, 1 practical). At the same time, the potential scores available for each assessment can affect the range of total scores achieved. The University of Manchester awards scores based on five grades (Unsatisfactory – Distinction), the University of Aberdeen awards scores on a scale of 9-20, whereas other universities award percentages. Retaining only graded data is a university-wide policy in at least five institutions.

Analysis of pilot EPM scores and application data for FP2010

Figure 3 illustrates the Pearson's Product-Moment Correlation Coefficient[5] between various pairs of measures relating to the performance of applicants to the Foundation Programme 2010. The first six rows compare measures that were collected during the live recruitment round, namely the original medical school quartile, the application form score

[5] Pearson's Product-Moment Correlation Coefficient is a measure of dependence between two variables. A correlation of +/-1 is a perfect (linear) correlation, and a correlation of 0 shows no correlation.

(excluding Question 1A and Question 1B), and the scores for Question 1A and Question 1B. The remaining rows include comparisons with and between the pilot EPM scores.

n.b. Application form score (60 points) – Q1A, Q1B and five 'white space' questions, mapped against the Foundation Programme person specification, with a total of ten points each
Question 1A (total 6 points) – previous degrees and intercalated degrees (national scoring criteria)
Question 1B (total 4 points) – prizes, presentations and publications (national scoring criteria)

Figure 3: Correlation for pairs of measures for applicants to FP2010 (22 schools)

Measures	Pearson's Correlation
Medical school quartile v application form score	0.18
Medical school quartile v Question 1A score	0.18
Medical school quartile v Question 1B score	0.12
Application form score v Question 1A score	0.08
Application form score v Question 1B score	0.04
Question 1A score v Question 1B score	0.24
Medical school quartile v written EPM score	0.83
Medical school quartile v practical EPM score	0.65
Application form score v written EPM score	0.15
Application form score v practical EPM score	0.16
Written EPM score v practical EPM score	0.51
Written EPM score v Question 1A score	0.18
Written EPM score v Question 1B score	0.11
Practical EPM score v Question 1A score	0.10
Practical EPM score v Question 1B score	0.08

Three pairs of measures indicate significant concurrent validity, namely the medical school quartile and written EPM score, medical school quartile and practical EPM score, and written EPM score and practical EPM score. The correlation with the practical EPM scores is slightly less than with written EPM scores, suggesting that schools may not have previously place as much emphasis on practical assessments when devising quartiles for recruitment to the Foundation Programme.

Analysis of medical school quartiles and EPM scores

Comparison of medical school quartiles and written EPM scores (22 schools)

- There is a correlation of 0.83, indicating a strong correlation between the assessments – and weightings – used by schools to inform quartiles as to inform written EPM scores
- 67% of students were in the same quartile for written EPM scores as with medical school quartiles. 30% of students moved to one quartile either side of their original rank, and 3% of students moved by more than 2 quartile ranks[6]

Comparison of medical school quartiles and practical EPM scores (20 schools)

- There is a correlation of 0.65, indicating a moderately strong correlation between the assessments – and weightings – used by schools to inform quartiles as to inform written EPM scores.

[6] i.e. applicants were moved from 1st to 3rd or 1st to 4th; 2nd to 4th; 3rd to 1st; 4th to 2nd or 4th to 1st

- 46% of students were in the same quartile. 44% of students moved to one quartile either side of their original rank, and 10% of students moved by more than 2 quartile ranks

An analysis of some of the correlations by medical school is presented in Figure 4 (the cells are blank where data was not available).

Figure 4: Correlation between medical school quartiles and EPM scores, by school

Medical school	Application form v original quartile	Medical school quartile v written EPM	Medical school quartile v practical EPM
Aberdeen	0.32	0.92	0.76
Barts and The London	0.26	0.93	0.48
Belfast	0.28		
Birmingham	0.37	0.95	0.49
Brighton and Sussex	0.17	0.82	0.53
Bristol	0.33	0.87	0.66
Cambridge	0.28		
Cardiff	0.36	0.83	0.67
Dundee	0.24		
Edinburgh	0.22	0.93	0.84
Glasgow	0.09	0.79	0.69
HYMS	-0.05	0.79	0.59
Imperial	0.21		
Keele	0.06	0.89	0.78
King's College London	0.34	0.80	0.68
Leeds	0.19	0.77	
Leicester	0.19	0.79	0.63
Liverpool	0.21	0.83	0.49
Manchester	0.20		
Newcastle	0.24	0.81	0.57
Nottingham	0.20	0.62	0.76
Oxford	0.12		
Peninsula	0.28		
Sheffield	0.35	0.90	0.75
Southampton	0.34	0.75	
St Georges, London	0.33	0.85	0.65
UCL	0.23	0.80	0.85
Uni. of East Anglia	0.15	0.76	0.78
Warwick	0.01	0.75	0.50

It is interesting to note the different levels of correlation between the two EPM scores and medical school quartiles. Only in one school – Sheffield – is the correlation stronger between medical school quartiles and the practical EPM scores, than with written EPM scores.

Feedback from medical schools in collating quartiles suggested that most of the difference in rank place relates to students at the margin. The EPM framework required schools to use the passmark for any resit assessments; however, subsequent feedback highlighted that many schools use the first-attempt mark. This created clustering at the

margins, as more applicants who had previously received their first-attempt mark, received the pass-mark. The piloted EPM framework also stipulated the weighting of assessments in the early and later years of the course. Where the weightings differed from those used by schools to inform quartiles, this also created movement in the ranked places to reflect student performance.

This analysis shows that the ranking a student achieves varies according to the ranking formula used. A common approach to producing EPM scores would reduce variability between schools whilst taking into account the approach to assessment and progression within schools.

Analysis of Question 1A and Question 1B

Data relating to additional points for previous degrees, and for prizes, publications and presentations, can be correlated with the pilot EPM scores using points awarded to applicants for Question 1A (previous degrees) and Question 1B (prizes, publications and presentations) on the white space application form.

Figure 4 illustrates a wide variation in the degree of correlation between medical school quartiles and application form scores, which have a weak positive correlation (0.22) overall. This suggests that the white space questions may be measuring different qualities and competencies of applicants, particularly for Question 1A and Question 1B. Depending on the results of pilots in 2010-11, the white space questions may be replaced with a Situational Judgement Test (SJT). Thus points awarded to recognise previous degrees, prizes, presentations and publications may only be incorporated within the EPM. The following analysis explores how these components are currently combined.

Question 1A (previous degrees – maximum 6 points)

- 57% of students received points for Q1A, of whom 34% were also in the 1st quartile
- 67% of students in the 1st quartile received points for Question 1A, compared with 48% in the 4th quartile
- Overall 7% of students scored between 1 and 4 points; 35% of applicants scored 5 points (55% of whom were in the 2nd and 3rd quartiles); 15% of students scored 6 points (76% of whom were in the 1st and 2nd quartiles)

Question 1B (prizes, publications, presentations – maximum 4 points)

- 25% of students received additional points for Q1B, of whom 34% were also in the 1st quartile
- 31% of students in the 1st quartile received additional points for Question 1B, compared with 18% in the 4th quartile
- Overall 17% of students received 1 point for Question 1B; 6% 2 points; 2% 3 points; and 0.4% 4 points

This analysis suggests that whilst the minority of applicants gain points for prizes, publications and presentations – and just over half gain points for previous degrees - the award of additional points does provide some additional discrimination at the upper end of the spectrum.

5. Combining scores for two quartiles

One of the aims of the EPM is to improve granularity from the current four quartile ranks[7]. By collecting two sets of quartiles of performance at medical school – written EPM scores and practical EPM scores – the two quartile scores could be combined in some way to achieve additional granularity. In principle, other bandings for example septiles or deciles, would achieve a higher level of granularity, although the feasibility and desirability of bandings at this level would need to be considered.

[7] Where more than one student achieves the same mark, or there is an odd numbered cohort size, it may not always be possible to divide a cohort into four equal groups.

Figure 5 illustrates the moderately positive correlation (0.51) between the written and practical EPM scores, which shows concurrent validity for the two measures. The cells in the table show the number of applicants with a given combination of practical and written EPM scores. 43% of applicants were placed in the same quartile for performance on written assessments (curriculum knowledge) and practical assessments (clinical skills).

Figure 5: Comparison of written EPM quartile and practical EPM quartile (23 schools)

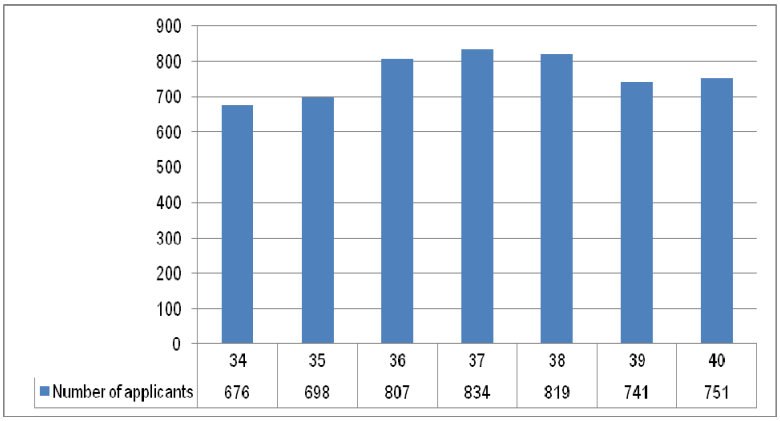
Number of applicants		Practical EPM quartile			
		1st	2nd	3rd	4th
Written EPM quartile	1st	751	407	202	89
	2nd	334	428	344	183
	3rd	189	338	451	371
	4th	63	173	327	676

One way of combining two quartiles (assuming a weighting of 50:50, as informed by the two consultations with medical schools) is to award points based on their quartile achieved, as illustrated in Figure 6. The points currently allocated for performance by quartiles is used for illustration. The distribution of students achieving each of the points (34-40) is illustrated in Figure 7, which illustrates additional discrimination at the upper and lower ends of the spectrum.

Figure 6: Points awarded for written EPM scores and practical EPM quartiles (combined by addition)

Points awarded		Practical EPM quartile			
		1st	2nd	3rd	4th
Written EPM quartile	1st	40	39	38	37
	2nd	39	38	37	36
	3rd	38	37	36	35
	4th	37	36	35	34

Figure 7: Points awarded for written EPM scores and practical EPM scores (combined)



Achieving 50:50 weighting between written EPM scores and practical EPM scores

Statistical advice regarding the method of combining two scores, and achieving a 50:50 weighting of the two elements, was sought. A greater degree of granularity can be achieved by combining two sets of quartiles, as illustrated in Figures 5, 6 and 7.

In order to achieve a true 50:50 weighting between two measures, dividing applicants into quartiles can only take place after the combination has been achieved. By dividing applicants into two (or more) quartiles before combining

the scores, the actual score (normalised) and position within the banding, is no longer taken into account. In Figure 8, applicant A is ranked 49th /100 (2nd quartile) and applicant B is ranked of 26th /100 (2nd quartile). However, when the quartiles are allocated before combining, both applicants achieve the same score.

Figure 8: Example of two applicants – scores and quartiles

	Written EPM rank/100	Written EPM score	Practical EPM rank/100	Practical EPM quartile	Written EPM score + Practical EPM quartile
Applicant A	49	2nd	49	2nd	4 (2+2)
Applicant B	1	1st	51st	3rd	4 (3+1)

Figures 9 and 10 illustrate alternative ways of combining scores based on two quartiles. Neither method takes into account the position within the band, or the underlying score achieved by the applicant, and can thus be rejected on grounds of unfairness. It is interesting to note that multiplying scores for quartile ranks provides greater differentiation between applicants in the 3rd and 4th quartiles. Division of quartiles distorts scores, depending on which EPM score is the denominator, but awards the same number of points for an applicant in the top quartile for both EPM scores as for an applicant in the bottom quartile for both EPM scores.

Figure 9: Points achieved by multiplying the points for quartile rank

Quartile		Practical EPM quartile			
		1st	2nd	3rd	4th
Written EPM quartile	1st	1	2	3	4
	2nd	2	4	6	8
	3rd	3	6	9	12
	4th	4	8	12	16

Figure 10: Points achieved by dividing the points for quartile rank

Quartile		Practical EPM quartile			
		1st	2nd	3rd	4th
Written EPM quartile	1st	1	0.5	0.33	0.25
	2nd	2	1	0.67	0.5
	3rd	3	1.5	1	0.75
	4th	4	2	1.33	1

Calculations to combine banded scores such as quartiles are less fair as they do not take into account the position within the band. It is not possible to achieve a true 50:50 weighting without taking into account the mark achieved in relation to the pass-mark and/or the cohort, and the difficulty of the exam. Any calculations based on raw scores would need to be normalised to take local factors into account.

Some schools adhere to a university-wide policy that only bandings or gradings are retained for student assessment performance. In this sense, the composite scores for the EPM already diverge from the specified module or assessment weightings. One way to remedy the issue of already graded data, or data with a small range between

scores, is to use a 'basket' of assessments. Whilst different assessments will be measuring different sets of skills, it can be assumed that with more assessments in the 'basket', the more closely converged the average score will be.

6. Changing the scoring system

Quartiles are based on the assumption that graduates from all medical schools are of equal calibre, such that a student in the 1st quartile from one school is of the equivalent standard of a student in the 1st quartile at a different school. Medical school assessments are quality assured by the General Medical Council, and the minimum standard of graduates from any school is assured to be of the same minimum standard.

The report of *Selection to Foundation: An Option Appraisal* deemed that a measure of performance at medical school could only be taken in relation to the cohort from that medical school. Owing to the timing of the application to the Foundation Programme, it is not possible to use Finals. There is no mechanism by which to standardise assessment scores from UK and non-UK medical schools to take into account the number of assessments, pass mark, level of difficulty – or to take into account that performance might be banded by individual assessments. Performance within cohort is the only fair, feasible and defensible measure of performance at medical school.

In the long term, development of assessment items in collaboration with all UK medical schools, through the Medical Schools Council Assessment Alliance (MSC-AA), will address issues of comparability between graduates from different medical schools. By sharing expertise in assessments, and by sharing assessment items themselves, it is anticipated that the collaboration will i) reassure the General Medical Council, NHS employers and the public that medical schools are looking in detail at the equivalency in standards between medical schools, ii) ensure confidence in the quality of medical school graduates, and iii) improve the student learning experience.

Distance from the mean vs distance from the median

One of the concerns about quartile bandings is regarding differentiation at the boundaries. By taking into account the distance from the mean, rather than median, it may be possible to more fairly reflect the performance of individual students in relation to their cohort, with the result that different bandings have a different number of students within. For illustration, the distribution of written EPM scores from Keele Medical School is illustrated in Figure 11, with a mean of 877.28 and a standard deviation of 77.72.

Figure 11: Distance from the mean, Keele written EPM scores

	Boundary score	Number of applicants
Mean -4 St.Dev	566.39	
		1
Mean -3 St.Dev	644.11	
		3
Mean -2 St.Dev	721.83	
		12
Mean -1 St.Dev	799.55	
		36
Mean	877.28	
		36
Mean +1 St.Dev	955.00	
		17
Mean+2 St.Dev	1032.72	
		0
Mean +3 St.Dev	1110.44	

Figure 11 illustrates differences in the number of individuals in a septile band, when taking into account difference from the mean as opposed to distance from the median. There are four applicants from Keele Medical School who scored significantly lower than other students in the cohort, but were allocated into the same quartile as students who were within two standard deviations of the mean. This would suggest that banding by standard deviations from the mean would achieve two advantages over quartiles, in that it would provide greater differentiation at the extremes, and could provide additional granularity – in this instance, six bandings compared with four quartiles. However, assuming a bell-curve of student performance, it would result in more tied ranks around the middle and thus less differentiation. Further, the calculations to devise rankings of performance in cohort would be more complex, with a greater risk of error, and less transparent to the non-expert

In order to compare distance from the mean in terms of standard deviations, the difficulty of the assessments would need to be known. Furthermore, it would need to be known that the distribution of scores followed a normal distribution. A goodness of fit test for a normal distribution, using the Chi-squared distribution, was performed on practical EPM scores from one school - Aberdeen - for illustration. However there was insufficient evidence to either accept or reject that the data (combined number of assessments) follow a normal distribution. Further tests would need to be undertaken to determine whether the total dataset, as well as its composites, follow a normal distribution.

It is extremely difficult to compare distance from the mean between cohorts, unless the data follows a normal distribution, and that scores have been standardised to take into account difficulty of assessment and pass mark. Whilst in one cohort, students may all perform within 1 or 2 standard deviations of the mean, in another there may be much greater variation. Similarly, in one school the mean may not be close to the median, such that a greater number of students may score above than the mean than below it.

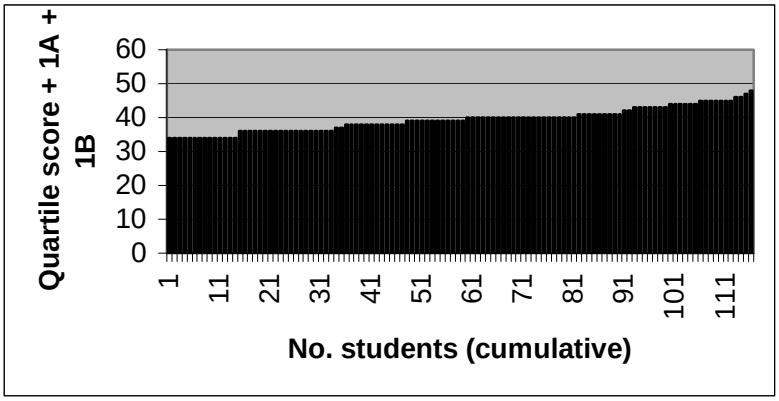
Statistically, distance from the median is a more robust measure, as unlike the mean, the median is unaffected by relatively high or low values. On the basis of this analysis, it is not possible to compare individuals in different cohorts using distance from the mean.

Increasing granularity (deciles, centiles)

One method of increasing granularity achieved by EPM scores is to divide scores into a higher number of bandings than quartiles, for example deciles or centiles.

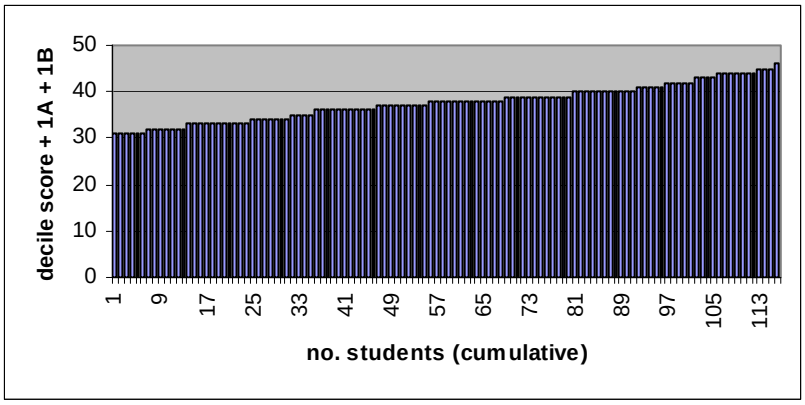
Figures 12 and 13 illustrate the effect of increasing the range of EPM scores, illustrated using data for practical EPM scores from Hull York Medical School. Points for quartiles (34, 36, 38, 40) and deciles (31, 32...39, 40) were combined with points for Question 1A and Question 1B (maximum of 10 points). Applicants under each scoring methodology could achieve a maximum of 50 points (1st quartile plus 10 points for Q1). 64 of 116 applicants scored 0 for both parts of Question 1.

Figure 12: EPM quartiles plus Question 1A and 1B, illustrative example (HYMS practical EPM quartiles)



Score achieved	34	36	37	38	39	40	41	42	43	44	45	46	47	48
Number of applicants	14	19	2	12	12	22	9	2	7	6	7	2	1	1

Figure 13: EPM quartiles plus Question 1A and 1B, illustrative example (HYMS practical EPM quartiles)



Score achieved	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
Number of applicants	6	7	11	7	4	11	9	13	12	11	5	5	4	7	3	1

Increasing the range of possible EPM scores achieved by applicants inevitably increases the range of possible total scores when combined with other performance measures, in this instance, educational achievements as evidenced on Questions 1A+B. It is interesting to note that although the full range for the deciles is 31-50 points, applicants scored between 31-46 points when their decile points score was combined with their score on Question 1 (range of 17). Similarly, when divided into quartiles, applicants could achieve 34-50 points, yet the scores achieved ranged from 34-48 (range of 15). With both scoring systems, there continue to be a number of applicants achieving tied scores across the range of scores.

With the example illustrated in Figures 12 and 13, the distribution of scores with quartiles produced a higher number of applicants with tied scores towards the lower end and middle of the scale (median 38 points), and a higher number of applicants with tied scores (e.g. 19, 22). In contrast, the distribution of scores with deciles produced a greater spread of applicants with different scores, with the most tied scores clustered in the middle of the scale (36-40 points, median 39 points) and fewer applicants with tied scores (11, 12,13).

Earlier analysis (Figure 3) indicated the small positive correlation between performance on EPM scores and performance on Question 1A and Question 1B. With both scoring systems, there is greater granularity at the top end of the scale, supporting the assumption that the scoring scheme identifies and rewards the highest achieving applicants.

In summary, increasing the number of bandings, such as from quartiles to deciles, can help to differentiate applicants by i) providing a greater spread of scores and ii) reducing the number of applicants with tied scores, without greatly affecting the values of scores achieved (mean, median, mode). Issues of feasibility and desirability need to be addressed.

Increasing granularity (changing the difference in points awarded to quartiles)

A second method of increasing granularity achieved by EPM scores is to allocate a greater difference in points for existing quartile – or other – bandings.

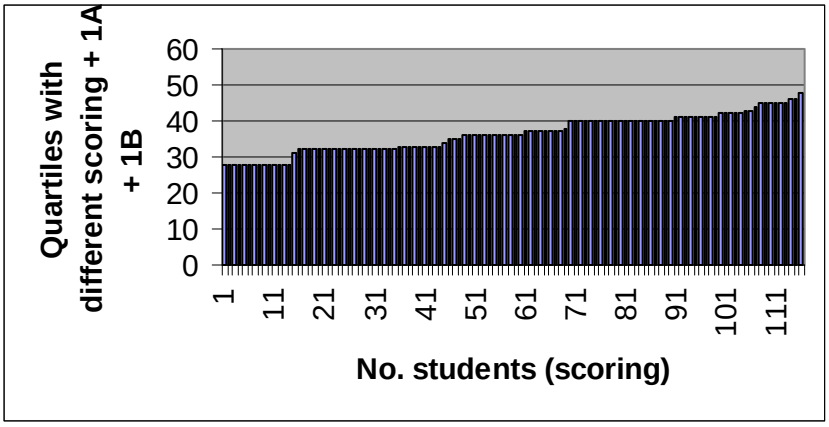
Figure 14 illustrates the effect of increasing the range of points available to quartiles, illustrated using data for practical EPM scores from Hull York Medical School. Points for quartiles (28, 32, 36, 40) were combined with points for Question 1A and Question 1B (maximum of 10 points).

It is interesting to note that although the full range for the deciles is 28-50 points, applicants scored between 28-48 points when their quartile points score was combined with their score on Question 1 (range of 21). A range of 17 possible scores were achieved by applicants within this range (not 29, 30, 39, 47).

Within this scoring system, there continues to be a number of applicants achieving tied scores across the range of scores, widely mirroring the range of scores illustrated in Figure 12 (combining quartiles plus Question 1A and 1B), with up to 20 applicants with the same tied score. The median of 36 reflects the lower range of points available to applicants.

In summary, increasing the difference in points awarded to bandings, such as quartiles, can help to differentiate applicants by i) providing a greater spread of scores and ii) reducing the number of applicants with tied scores. However the additional level of granularity is comparatively less than that achieved by increasing the number of possible bandings. The desirability of changing the range of points allocated to quartiles need to be addressed.

Figure 14: Quartiles (alternative scoring) plus Question 1A and 1B, illustrative example (HYMS practical EPM quartiles)



Score achieved	28	31	32	33	34	35	36	37	38	40	41	42	43	44	45	46	48
Number of applicants	14	1	20	9	1	3	12	8	1	21	9	5	2	1	6	2	1

7. Other findings

The pilot generated a considerable amount of helpful feedback from medical schools. The feedback, summarised below, is set out in greater detail in Appendix B.

Practical versus written assessments. It is not possible for all schools to provide separate scores or rankings for practical and written work, as some schools have assessments that combine both competencies.

Timing of course delivery and assessments. There is considerable variation in the timing of courses and assessments. Different cohorts – often within the same school – may be at medical school for four, five or six years. It is challenging for schools to standardise the assessment data for students who have not taken the same assessments, or the same combination of assessments. Given this, the EPM framework should not try to specify that certain assessments should be taken from certain years to produce an EPM.

Grading and banding. There is considerable variation in the way in which the outcomes of assessments are quantified and recorded by universities. Some universities have policies that require schools to keep raw scores from assessments; others require schools to retain only broad gradings. Some assessments yield very granular results; at the other extreme some assessments result only in a pass or a fail. There are also more complicated scoring schemes including merit and demerit points, negative marking and where the number of passes or fails of individual questions determine an overall pass or fail. The EPM framework needs to cater for this degree of variation.

Transfers, Student Selected Components etc. There is a particular difficulty in finding a satisfactory way to deal with students who transfer between schools, as such students will not have undertaken the same mix of assessments as the others in the school to which they have transferred. Schools currently take one of two approaches: a complex process of standardisation, or developing quartiles according to assessments that all students have taken in common. There is a similar problem in dealing with other cases in which not all the students at a school undertake the same assessments, for example where the students have some freedom to select modules of work.

Re-sits. There needs to be an agreed standard way of allocating marks in relation to re-sits, taking account of any extenuating circumstances, and the varying opportunities to re-sit failed finals in the same academic year.

Authorised absence. There needs to be an agreement about how to treat cases in which a student is authorised to miss an assessment that is taken by the rest of the cohort.

Comparability across schools. The basis of the EPM (and the original quartile score) is an assumption that the cohorts are broadly comparable across medical schools. Some schools have expressed concerns that this assumption may not be valid.

8. The way forward

The EPM pilot has demonstrated the importance – and the benefits – of a nationally agreed framework for EPM scores in terms of transparency and consistency. The 25 medical schools which participated fully in the EPM pilot were able to return raw data relating for both written EPM scores and practical EPM scores. Feedback from the remaining five schools indicated that they too would be able to provide data on curriculum knowledge and clinical skills, although they were unable to adhere to the framework within the timeframe.

Whilst many, if not all, schools consult on and publish the composition of quartiles, it is evident that there is wide variation between medical schools in the way they produce quartiles. The pilot revealed the different compositions of current measures of performance at medical school – which include assessments in all years in some schools, and a single year in others. It is not practical to specify the exact composition of the EPM in terms of specific assessments, as the different approaches to assessment and learning should be encouraged. Nationally agreed EPM standards would be beneficial in avoiding unnecessary variation across schools, for example by including a requirement for a standard way of allocating scores to students who re-sit assessments. The movement of students between medical school quartiles and EPM scores further highlights the importance of a defensible and consistent approach to developing EPM scores.

The EPM pilot, and the issues raised by schools in adhering to the EPM framework, highlighted that a framework which specifies X% of assessments from one year and Y% from another is inappropriate, since it does not recognise the variation across schools in the timing of courses and assessments. Similarly a framework that calls for a clear split between assessments of curriculum knowledge and clinical skills is not feasible, as one competency underpins the other, and assessments frequently assess both competencies.

In principle there is scope to make the EPM more granular than quartiles, provided that the underlying data enable sufficient differentiation between students. This granularity is separate to the weighting that might be used for the EPM with the SJT. It may be prudent for there to be a requirement for a minimum number of assessments to be taken into account, to ensure that the desired granularity can be reached.

Averaging a student's performance over a 'basket' of assessments should be more representative of the student's capability than taking any one of the assessments individually. The pilot has shown that the rules for constructing such a basket cannot be specific about the nature or timing of the assessments to be taken into account, or about the split between written and practical assessments, as these factors vary by school. The EPM framework needs to be defined in a sufficiently generic way to take this variation into account.

Given the above, a pragmatic approach would be as follows:

Each medical school should calculate an overall average score for each of its students, based on a basket of assessments of clinical skills and curriculum knowledge. Each school will be responsible for constructing the most appropriate basket of assessments for the purpose of providing the EPM, taking into account local factors. The basket should not be artificially skewed. For transparency, all schools should develop the composition of EPM scores in consultation with its students, and would be required to publish the details of the basket to its students.

The students in a cohort should be competition-ranked according to their score, then given an overall EPM at an agreed level of granularity (e.g. quartiles, deciles, etc).

To ensure that EPM scores are determined in a consistent and fair way across schools, there must be standards to address:

- General principles to be applied in calculating the weighted averages of assessments
- A minimum number of assessments
- Re-sits (either of single assessments or of an entire year)
- Missing assessment scores (e.g. where an applicant has been on sick leave during an assessment)
- Treatment of pass/fail assessments and other forms of banding
- Transfer students
- Cases in which different students within a cohort take different assessments
- Cases in which the strict boundary of a class (e.g. quartile) falls among a number of equally-ranked students
- Whether and how additional points may be awarded for other achievements (e.g. prizes, publications, presentations and extra degrees)

A draft of these standards should be worked-up by an EPM Task & Finish Group, reporting to the ISFP Project Group, through liaison with those schools who are particularly affected by the issues. The revised EPM standards should then be issued for consideration by all UK medical schools and the ISFP Project Group. Legal opinion will be sought.

Appendix A: Draft Framework for Educational Performance Measurement (EPM)

Component 1: Written assessments – 50% of EPM

This component uses the results of summative written assessments throughout the student's undergraduate career, with the assessments in the later years more heavily weighted. Please determine the total mark for each student using the weighting below.

- Early years (40% of Component 1)
    - Year 1 of 4 year graduate entry courses
    - Years 1 and 2 of 5 year courses and
    - Years 1-3 of 6 year courses*
- Later years (60% of Component 1)
    - Years 2 and 3 of 4 year graduate entry courses
    - Years 3 and 4 of 5 year courses and
    - Years 4 and 5 of 6 year courses*

Component 2: Practical assessments (e.g. OSCEs) - 50% of EPM

This component uses the results of summative practical assessments, such as OSCEs. Please determine the total mark for each student using the weighting below.

- Years 2 and 3 of 4 year graduate entry courses
- Years 3 and 4 of 5 year courses and
- Years 4 and 5 of 6 year courses*

1. Data will be drawn from summative rather than formative assessments.
2. Schools will need to devise their own weightings for modules and assessments – which may include summative SSCs - within the three components above. These weightings will be made transparent.
3. Where re-sits have been taken, the marks awarded will be capped at the pass mark. An exception to this will be in the case of extenuating circumstances, as approved by an examinations board or committee, for which the marks from the repeat assessment will then be used.
4. Schools whose students move after 3 years will provide the receiving School with the necessary data to allow early years' results to be included in the calculation.
5. The calculations will be made using actual marks (standard-set) awarded for modules and not the grade points, to keep as much discrimination as possible.

*Six year courses including a Foundation or pre-clinical year should count the foundation year as Year 0, with the following five years as Years 1-5.

Please note that the length of the course excludes time out for intercalation.

Appendix B: Issues raised by medical schools in adhering to the EPM framework

Difficulties in accessing data

- Difficult to separate scores for written and practical competencies, as assessments usually include both and it is difficult to assess the appropriate weightings to be applied retrospectively (Barts, Birmingham, Cambridge, Edinburgh)
- Data are held by another school within the university (Cardiff – years 1&2 are managed by the School of Biosciences)
- Data for early years assessments is not held for transfer students (Aberdeen, Barts, Cardiff, Edinburgh, Imperial, KCL, Manchester, Oxford, Imperial, UCL)
- Data is not stored in a format that enables the framework to be followed easily – many schools reported that it was a manual process, for example UCL referred to 36 different spreadsheets plus more to track students who didn't progress normally, and checking individual cases of extenuating circumstances.

Granularity

- University-wide policy to record assessment bandings or gradings, not raw scores (Aberdeen, Leeds, Liverpool, Manchester, UEA, Warwick)
- Some assessments are pass/fail (Liverpool, Southampton, UEA)

Ranking within a cohort

- Students from different intakes (i.e. graduate entry, five year and six year courses) ranked separately (Imperial, St George's)
- Students from different intakes ranked together (KCL, Southampton)

Normalisation

- Students from the same intake do not take the same combination of assessments (Manchester)
- Students from different intakes (four, five and six year courses, plus direct entrants to Year 2 in Aberdeen, to Year 3 for BDS graduates in KCL) do not take the same combination of assessments (Aberdeen, Birmingham, HYMS, Imperial, KCL, Manchester, Southampton, UCL). This is currently addressed by two methods – using only data common to all intakes (KCL – 3 years; Manchester – 2 years; Southampton – 1 year), or using a complex process of standardisation for the pass mark, level of difficulty and scale (e.g. Aberdeen)

Transfer students

- Transfers of whole cohorts (St Andrews to Manchester, Swansea to Cardiff – both at year 3)
- Transfers of small groups of students – from Oxbridge, from the International Malaysia University (Aberdeen, Barts, Cardiff, Edinburgh, Imperial, KCL, Manchester, Oxford, Imperial, UCL)

Composition of EPM scores

- The framework limits the number of assessments to be included
    - Clinical skills assessments in early years are excluded (BSMS, Brighton, Birmingham, Cardiff, Edinburgh, Oxford, Warwick, Sheffield)
    - Assessments with pass/fail are excluded from the framework if resit marks are capped at the passmark (BSMS usually includes 49 assessments within quartiles, but the EPM framework is limited to 14)
    - Cambridge places a greater emphasis on practical assessments – only two assessments are purely 'written'
- The framework requires additional assessments to be included

- Interpretation of definitions of 'practical' and 'written' – i.e. mode of assessment vs competencies being assessed
    - Cardiff - assessments termed 'skills and competencies' include academic skills panel, group work, critical reading, essay under exam conditions
    - St George's – includes Student Select Components
    - Edinburgh – includes Professionalism/ PPD measures, psychiatry viva, interpretations of x-rays assessed online, mini-CEX
    - Sheffield – moving towards greater emphasis on workplace-based assessments, only one OSCE (year 3)
    - UEA – includes presentations

Resits

- Use of actual first mark (Aberdeen, Birmingham, Liverpool, Manchester, Sheffield, St George's, UCL) vs capped at pass mark (Imperial, Sheffield) vs compensated fail (Edinburgh, St George's, UCL)
- Verifying extenuating circumstances in retrospect is a time-consuming and manual process
- Capping resits at the passmark leads to clustering at the lower end of student performance

Other issues for agreement

- Is the divide between written and practical artificial? Definition of 'written' and 'practical' – in terms of competencies?
- Enable flexibility for Schools to include other assessments (e.g. SSMs, professionalism measures, portfolios) etc – but in a way that does not contract the goal of standardisation across schools
- Timeframe to introduce framework - not fair to impose a framework retrospectively – if early years data are to count, this must be openly agreed from the start of the course
- What of authorised absence and no mark returned? In Manchester, the mean mark for the module is used.
- Inclusion and weighting of points for presentations, publications, prizes or previous degrees

Additional comments

- Aberdeen - Putting in a pass mark of 9 for any students who resit a course means that the bottom quartile will be more clustered. This was something that our students felt strongly about; the students who have been at the lower end of the class but who have never failed an exam are then clustered in with students who have re-sat, sometimes across lots of exams, and they felt that this was not fair.

- Barts - Weighting of pre clinical and clinical years. As we do not differentiate significantly in terms of content this is a fairly artificial division for us and no doubt some other schools. Early and late years are a better distinction, but then the weighting is for the value of the contribution to the finishes product rather than the proportion of time spent in a clinical setting.

- Edinburgh - Having been through the process of allocating marks to either 'Written' or 'Practical' and faced the difficulties we have highlighted, we wonder if the effort of splitting marks in this way is worthwhile, especially in the absence of an underpinning principle or construct for the division. I understand UKCAT may also be moving to considering 'clinical' versus 'academic' abilities, another difficult distinction to make in practice since academic knowledge underpins most clinical activity and practical clinical activities such as interpreting data and x-rays, and even prescribing might be assessed by way of an online 'written' exam.

  Perhaps competencies, such as history taking, information giving, and clinical examination, demonstrated through interactions with patients, are the important ones to distinguish, since in practice there appears to be a handful of students who, though able in other ways, struggle to master these competencies. However these

are most robustly tested in formal summative tests such as OSCEs to overcome the leniency of workplace-based assessments and this would result in very few assessments that could be collated for such a mark.

- Leicester – the (ranking) order will be very different in the middle, due to the different weightings, although the students at the top and bottom should be the same. Pass/fail is a set standard– students have to pass minimum number of questions, not a minimum percentage – not a compensatory system.

- Oxford - I am afraid that I continue to regard this exercise as a red herring. Increasing the 'granularity' used for compiling scores to discriminate between students at individual schools is not going to improve the fairness of the process one jot -the problem continues to be the underlying assumption that EPM scores are comparable between medical schools, which the evidence suggests they are not. I have no problem with following a common methodology, but I do hope people will stop pretending that doing so is going to make selection fairer. The assumption that quartiles are equal between medical schools is of course made in the existing rankings, but the impact is low because so few points are awarded for the quartile rankings.

  If the EPM were used to increase the fraction of the ranking score allocated for medical school performance, we believe that our lower scoring students would be unfairly discriminated against, and consequently could not agree to it.

- Southampton – to ensure transparency, students are provided with their grades which they can then verify (based on single assessment). If the framework required complex normalisation this transparency would not be possible.

- St George's - We are in the midst of a change of curriculum and even aside from that, we had a big change to the structure of our pre clinical and clinical exams. So for the students on the list, anyone who did not go straight through the system and took a year out or intercalated etc will have done a different style of exam than other students.

  There is also a problem when students don't sit the exams with the year they are being ranked with i.e. a student who failed a Year 3 exam in 06-07 gets the pass mark for that exam for that year, but they are ranked with other students who failed a Year 3 exam a year later in 07-08. This student would get the pass mark for that exam, but if the pass marks are different each year (as they are every year), this could skew the ranking. The only way to get round this would be to % each exam and pass mark and make everything out of 100. This was too much work for the purposes of this exercise and would have taken me weeks and weeks to work out individual pass mark % for each student across all the exams of all the years, especially breaking up a combined exam into OSCE and written components.

  We have some exams that you can't pass or fail, and they just add into another exam which you can pass or fail overall. It is not formative per se as the overall exam, which includes the results from this previous exam as above, is summative. Therefore if someone gets below the pass mark for the first part, should I give them the pass mark or their actual mark? I gave them the pass mark, but they haven't actually failed this exam and do not resit it.

- Warwick - We do not use overall percentages in written examinations as a way of determining who has passed. Each question has a separate threshold that the student has to pass to be given a single point. These points are then added up and students have to reach a required level of points to pass. Points are also used to allocate grades. We therefore worked out our percentage for the written components based on

how many grade points out of the total available for example 36 points available 18 points achieved equalled 50%.

For our clinical examinations students achieve possible marks for certain grades whilst other are considered to be 'demerit' points. For these examinations we looked at whether students had been required to sit further tests or not and how they had done overall. 100% meant no demerit points, 75% demerit accrued but no further exam, 50% demerit points, further exam and passed, 0% demerit points and failed.

The results have moved students within quartiles and in one case a student has moved ranking from 131st to 27th! We are not happy with this as this student has now sat Finals and has been required to re-sit in November 2010. The additional information that we used in our original calculations and the instruction to put all re-sit candidates in the 4th quartile meant that this student was placed, in our opinion, in the correct quartile using the whole body of knowledge.

- UCL - The data requirements included many complexities that we considered and rejected at a series of working parties with representatives from Oxford, Cambridge and the other London Medical Schools when quartile rankings were first introduced and when we were first determining how to calculate quartiles for UCL students.

  We have a complication at UCL relating to IBSc's, which is that we currently have a compulsory IBSc for non-graduates which may be taken between years 3 and 4 or between years 4 and 5. It makes compiling the marks for UCL students complicated because we have to draw on exams taken in a number of different sessions by different cohorts and with different exams and different standard set pass marks. For the EPM pilot, because of the different cohorts and the structure of our summative exams, we had to work with 36 spreadsheets to compile data for students who had progressed normally, and more to track students who did not progress normally. For MBPhD students and others with long period of interruption we had to track results back to spreadsheets dating over 8-9 years. Each year had different standard set pass marks with, sometimes, considerable variation.

  The compilation of exam results dating back, including MBPhDs, for 8-9 years and matching it to the EPM cohort took perhaps a day of admin time. That's feasible and represents about twice the amount of time currently needed to compile our rankings based on a customised export of year 3 and year 4 exam results. However, in our current system for calculating rankings we sought deliberately to minimise manual manipulation of data because of the risk of error, and this is my main concern with the complexity of the pilot.

  Tracking extenuating circumstances posed the biggest problem and we also have concerns about the fairness of using extenuating circumstances because at UCL we apply a tariff depending on the severity of the extenuating circumstances and the likely impact on student performance. For candidates graded 1 for extenuating circumstances, the mitigation is very weak and we wouldn't feel comfortable allowing an advantage to these candidates by using a re-sit mark achieved after an additional year's study. Also, 5 years ago information about extenuating circumstances and tariffs awarded wasn't recorded reliably in our exam mark sheets and so we had to go back to different spreadsheets and to student files and paper records to work this out. It's very unlikely that this type of information will ever be stored in the electronic student database against exam results and so there will always be a risk that some candidates with extenuating circumstances will be missed when tracking cohorts of 380-400 students over 8-9 years of a programme, and that some candidates with weak mitigation will be advantaged if their re-sit score is used.

  Data manipulation increases the risk of error, and, even with easier compilation of exam marks, the amount of data manipulation in the pilot for re-sits and capping and extenuating circumstances is considerable. Even

with a proper database it would be impossible to put systems in place retrospectively to track capped/original marks and extenuating circumstances. In years 3 and 4, our compensation rules for calculating the final result mean that capping the written and clinical marks separately at the standard set pass mark for each components as requested for the EPM meant, in some cases, raising a re-sit mark for one component from below the pass mark and awarding a higher mark than was achieved at the re-sit attempt.

In summary, I think our main concern is that the amount of manual manipulation of data and the amount of cross-checking against different mark sheets with different standard examinations and different pass marks means that we could not be confident in using the data provided for the EPM pilot to calculate the real quartile rankings to be used for our students, and a simpler method would be preferable as it would reduce the risk of error and produce more reliable data for rankings.

Appendix C: Caveats for interpretation of EPM scores

Aberdeen

- Scores based on points awarded by band, not raw scores
- First attempt marks used, not capped at the pass mark

Barts and The London

- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks

Birmingham

- First attempt marks used, not capped at the pass mark

Cambridge

- The 'early years' part of component 1 includes both written and practical – it is not possible to split the marks retrospectively

Edinburgh

- The 'early years' part of component 1 includes both written and practical – it is not possible to split the marks retrospectively
- Practical scores include Year 2 OSCE
- First attempt marks used, not capped at the pass mark
- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks

Leeds

- Scores based on points awarded by band, not raw scores.
- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks
- First attempt marks used, not capped at the pass mark
- Practical assessments – based on Year 3 OSCE only

Liverpool

- Scores based on points awarded by band, not raw scores
- First attempt marks used, not capped at the pass mark

Manchester

- Scores based on points awarded by band, not raw scores.
- First attempt marks used, not capped at the pass mark
- Only data from Years 3 and 4 is used

Oxford

- Clinical exams are scored 1-4 (can be decimal places) and converted to percentages.
- First attempt marks used, not capped at the pass mark
- Scores for preclinical years calculated using a merit system and not raw data (3 points for distinction, 1 point for merit, failed paper cancels out points for merits)

- Scores for preclinical years includes marks from Yrs 1&2 for the graduate entry cohort, as science based written exams are in both years not just year 1

Sheffield

- First attempt marks used, not capped at the pass mark

Southampton

- Practical assessments – based on Year 3 OSCE only
- Written assessments – based on two Year 3 assessments only

St George's, University of London

- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks

UCL

- Transfer students – assumption that the mark achieved in years at that medical school used for early years marks

University of East Anglia

- Some modules are pass/fail (50/0 points) – however as the resit mark was capped at the pass mark, these modules were then excluded from the EPM pilot

Warwick

- Points are allocated to grades (up to 36 points)
- Clinical examinations are awarded 0%, 50%, 75%, 100% based on demerit points
- Only two clinical assessments could be used, as the framework did not allow for Year 1