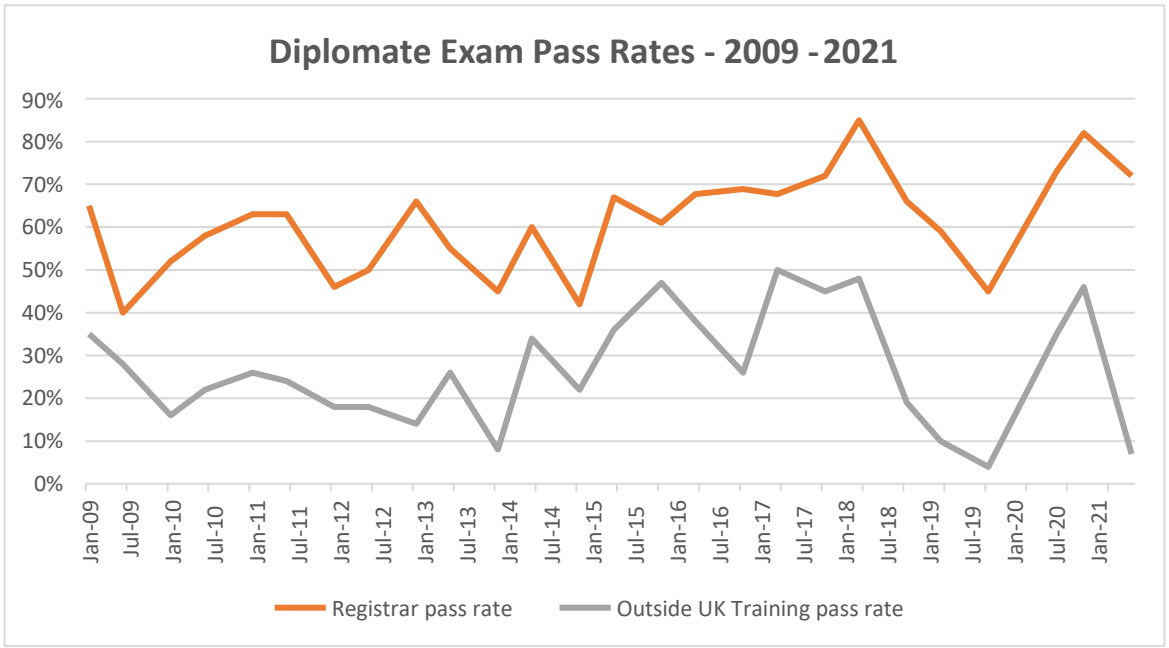
Diplomate Exam monitoring and performance

Background description

In parallel with the introduction of modified Angoff standard setting of our Diplomate examination in January 2017, the Diplomate exam team also introduced robust exam review by an independent external educationalist at each diet (sit). This has allowed the Chair/deputy Chair and Board to understand at an exam and question level, how the examination is performing at that diet, which is over and above simple monitoring of pass/fail rates.

Key variables that are monitored at an examination level are:

1. Exam reliability statistics – notably the Cronbach alpha. Values can range from 0-1. Higher values are better. Many written exams have Cronbach alpha values between 0.6-0.8. The target value for Cronbach alpha is 0.8 or above.
2. Generalisability statistics – these are another form of reliability statistic, but is considered a 'better' measure of reliability as it is less affected by outliers whose values can artificially elevate Cronbach alpha values.
3. Standard Error of Measurement (SEM): this is another measure of an examination's performance. A larger Standard Error of Measurement implies less certainty in the estimate of a candidate's true performance. Values can range from 0 upwards. Lower values are better (implying more accurate estimate of a candidate's performance). Examinations should aim for an SEM below 3.0. [Note: this is not the same as the Standard error of the mean]



The figure above previously indicated a rising trend in pass rates. However, closer inspection now

suggests two periods. The first period is exam diets preceding formal standard setting's introduction in January 2017. Prior to that point pass rates for those in UK training varied reasonably widely (from 40-70%) with an increasing pass rate trend observed. Subsequently, from January 2017 pass rates have continued to vary (indeed this is very noticeable in January 2020 when noise disruption significantly affected candidate performance). However, there is now no longer a rising trend observed in pass rates, instead these have varied between about 65 and 80% for those in UK training.

In terms of variability in pass rates, it remains important to be aware that modified Angoff standard setting does not of itself remove this. Instead, a number of factors including candidate variability will contribute. Furthermore, unlike large cohort undergraduate medical examinations, this exam is currently taken by approximately 60-90 candidates, which inevitably increases the variability observed above.

As was noted in our first report, modified Angoff standard setting is designed to ensure that we now explicitly set the pass mark for each and every question according to our expert view of the difficulty of a question, assessing what we believe a borderline competent candidate would score on each question. In line with best practice, our modified Angoff panels include a minimum of eight experienced examiners and commonly more. All examiners are in senior public health roles, in a wide variety of service and academic settings, and come from across the UK, with one-two representatives from Hong Kong.

Table 1: Paper and exam reliability statistics:

	Four diet rolling average					
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**		**Long-run average (Jan 17 to Oct 21**
Paper I Cronbach alpha	0.86	0.85	0.84	0.84		0.84
Paper I G-coefficient	0.85	0.84	0.83	0.83		0.82
Paper II Cronbach alpha	0.76	0.74	0.75	0.76		0.74
Paper II G coefficient	0.74	0.72	0.72	0.72		0.72
Exam reliability (Cronbach alpha)	0.90	0.89	0.89	0.88		**0.88**

Our target for an individual paper is a reliability value > 0.6, and for the exam ≥0.8.

Table 2: Standard Error of Measurement for papers and exam:

	Four diet rolling average					
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**		**Long-run average (Jan 17 to Oct 21**
SEM Paper 1	3.62	3.78	3.95	3.97		3.35
SEM Paper 2	3.88	4.05	4.40	4.63		3.69
SEM Exam	2.76	2.86	3.06	3.41		**3.09**

Our target for an individual paper is an SEM value < 4.0, and our target for the exam is ≤ 3.0

In terms of examiner performance, the key summary variable is each pair of examiner's intra-class correlation coefficient. Good alignment is shown with coefficients in excess of 0.6, and excellent where coefficients are over 0.75. The figures below indicate excellent alignment in scoring between our examiner pairs.

Table 3: Examiner performance – intra-class correlation (note single marking occurred in Mar 21, so no 4-diet average provided):

	Four diet rolling average					
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**		**Long-run average (Jan 17 to Oct 21**
Average ICC across all examiner pairs	0.84	0.83	-	0.83		0.84

In addition, we monitor the mark correlation before and after agreement for each question. Mark correlations reflect the quality of the questions we set, the mark scheme guidance, and examiner performance. Again, the data show good to very good correlation both before, and particularly after mark agreement.

Table 4: Examiner question correlations averaged across all questions before and after agreement:

	Four diet rolling average					
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**		**Long-run average (Jan 17 to Oct 21**
Paper I mark correlations before agreement	0.60875	0.6235	-	0.623		0.58
Paper I mark correlations after agreement	0.79425	0.801675	-	0.802567		0.78
Paper II mark correlations before agreement	0.642465	0.652382	-	0.614287		0.60
Paper II mark correlations after agreement	0.779048	0.759103	-	0.761767		0.75

Our target correlation is >0.5 before agreement and >0.7 after agreement.

Question-level performance: in addition, each Exam Board reviews detailed psychometric and performance data on all questions set. Two question indicators are reported as an overall summary of question performance:

- Facility: this reflects how easy or hard a question is. The % facility equates to the % of candidates who pass a given question. The data below show, in general, Paper I questions have higher facility than Paper II questions, and this remains a fairly static feature of these two papers. Questions in Paper I on average having a facility around 70-75%, and Paper IIB around 55%. Paper IIA has shown some relative variability, ranging from 42-58%.

Table 5: Facility (i.e. % candidates passing individual questions) **averaged across all questions:**

	Four diet rolling average					
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**		**Long-run average (Jan 17 to Oct 21**
Paper I facility	75%	72%	72%	70%		73%
Paper IIA facility	42%	42%	54%	58%		54%
Paper IIB facility	53%	55%	55%	57%		57%
Overall facility of examination	62%	61%	64%	64%		65%

No target set

- Discrimination: this reflects whether a question can distinguish between passing and failing candidates (overall). We use 27% discrimination. This measure compares the % of candidates passing the question amongst the top 27% of candidates and bottom 27% of candidates. A highly discriminating question would be passed by all of the top 27% of candidates, and failed by all of the bottom 27%. However, this measure needs to be interpreted with reference to question facility, as an 'easy' question which is passed by most candidates will automatically have a poor discrimination. The question may nevertheless be valid and useful. Scores range from -1 to +1, with higher scores indicating better discrimination. Our target discrimination is >0. Any question with a negative discrimination would be rigorously reviewed and is likely to be removed.

  The data below indicate excellent discrimination, which on average, appears to be improving with time.

Table 6: Question discrimination averaged across all questions

	Four diet rolling average					
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**		**Long-run average (Jan 17 to Oct 21**
Paper I discrimination	0.69	0.75	0.75	0.78		0.66
Paper II discrimination	0.87	0.87	0.89	0.89		0.81
Overall average discrimination	0.78	0.81	0.82	0.83		0.73

Our target for discrimination at paper level is >0.5

Chairs' summary 2017-19:

The above monitoring indicates that the exam continues to perform well psychometrically. We have consistently excellent reliability statistics. Our SEM has been creeping up over the last four diets, but remains acceptable, if slightly above (on average) our target of <3.0 for the exam as a whole. Our examiners continue to have consistently excellent intra-class correlation coefficients, and good correlation at question level before and after agreement. Our questions have always shown good discrimination, but this review notes that these (already high) levels of discrimination seem to be rising even higher, with values considerably in excess of our target of 0.5.

Overall, this is a very reassuring set of data, but we need to remain cognisant of our rising SEM. No immediate action is required, but this does need continued, close monitoring.