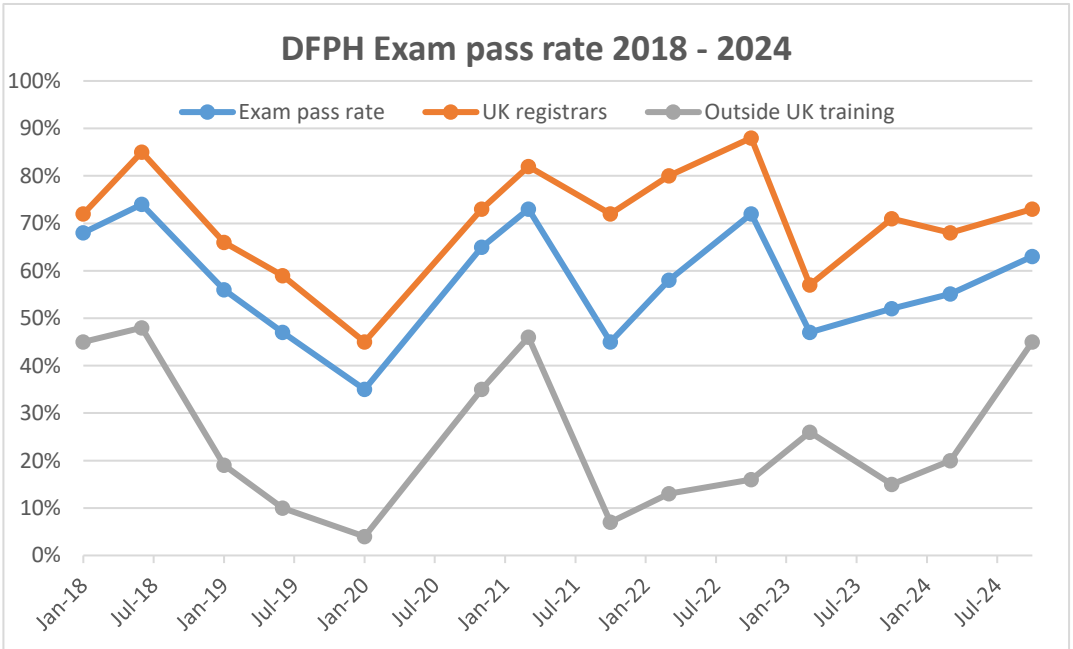
Diplomate Exam monitoring and performance

Background description

In parallel with the introduction of modified Angoff standard setting for our Diploma of the Faculty of Public Health (DFPH) examination (introduced in January 2017), the exam team also introduced robust exam review by an independent external educationalist. This has allowed the examiners and Board to understand how the examination is performing at each diet/sitting (at an examination, individual paper, and individual question level), over and above simple monitoring of pass/fail rates.

Key variables that are monitored at the examination and individual paper level are:

1. Exam reliability statistics – notably the Cronbach alpha. Values can range from 0-1, with higher values reflecting improved reliability. Many written exams have Cronbach alpha values between 0.6-0.8, however the target value for a high-stakes exam is 0.8 or above.
2. Generalisability statistics – these are another form of reliability measure but may be considered a 'better' measure of reliability as they are 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 related to the spread of observed scores and the exam reliability. A larger SEM 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 shows a declining pass rate in exam diets occurring immediately after the introduction of formal standard setting in January 2017, reaching a low point of below 50% in January 2020 (the last diet delivered in person as a written exam). Following the move to an online platform during the Covid-19 pandemic, pass rates for all candidates rose rapidly and were maintained at a high level for those in UK training until October 2023, where they peaked at almost 90% (for this group) and over 70% for all candidates. Subsequently, pass rates fell to just below 60% for those in UK training and below 50% overall, following which there has been a further upward trend for all candidates to October 2024.

In terms of the range of pass rates observed, it is important to be aware that the use of standard setting (modified Angoff approach) does not in and of itself remove variation in pass rates between exam diets. A number of factors will contribute to this, including candidate variability and candidate familiarity with the sections of the syllabus sampled at each sitting. Furthermore, unlike large cohort undergraduate medical examinations, this exam is currently only taken by approximately 80-90 candidates at each sitting, which itself inevitably increases the variability observed.

As was noted in earlier reports, 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 each question, assessing what we believe a borderline competent candidate would score on each question (or sub-question, where relevant). In line with best practice, our modified Angoff panels include a minimum of eight experienced examiners and commonly many more. All examiners are in senior public health roles in a wide variety of service and academic settings, and come from across the UK (all four nations), and include a representative from the Hong Kong College of Community Medicine.

Table 1: Paper and exam reliability statistics:

	Mean value						
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**	**Mar 22 to Oct 24**		**Long-run average (Jan 17 to Oct 24)**
Paper I Cronbach alpha	0.86	0.85	0.84	0.84	0.90		0.86
Paper I G-coefficient	0.85	0.84	0.83	0.83	0.89		0.85
Paper II Cronbach alpha	0.76	0.74	0.75	0.76	0.87		0.78
Paper II G coefficient	0.74	0.72	0.72	0.72	0.85		0.75
Exam reliability (Cronbach alpha)	0.90	0.89	0.89	0.88	0.94		**0.90**

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:

	Mean value						
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**	**Mar 22 to Oct 24**		**Long-run average (Jan 17 to Oct 24)**
SEM Paper 1	3.62	3.78	3.95	3.97	4.01		3.87
SEM Paper 2	3.88	4.05	4.40	4.63	4.69		4.33
SEM Exam	2.76	2.86	3.06	3.41	3.13		**3.04**

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 average is provided for this period):

	Mean value						
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**	**Mar 22 to Oct 24**		**Long-run average (Jan 17 to Oct 24)**
Average ICC across all examiner pairs	0.84	0.83	-	0.83	0.90		**0.85**

Note that single marking occurred in March 2021, so no average is provided for this period

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:

	Mean value						
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**	**Mar 22 to Oct 24**		**Long-run average (Jan 17 to Oct 24)**
Paper I mark correlations before agreement	0.61	0.62	-	0.62	0.74		0.65
Paper I mark correlations after agreement	0.79	0.80	-	0.80	0.86		0.81
Paper II mark correlations before agreement	0.64	0.65	-	0.61	0.71		0.65
Paper II mark correlations after agreement	0.78	0.76	-	0.76	0.85		0.79

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:**

	Mean value						
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**	**Mar 22 to Oct 24**		**Long-run average (Jan 17 to Oct 24)**
Paper I facility	75%	72%	72%	70%	66%		71%
Paper IIA facility	42%	42%	54%	58%	71%		53%
Paper IIB facility	53%	55%	55%	57%	64%		57%
Overall facility of examination	62%	61%	64%	64%	67%		**64%**

No target set

- Discrimination: this reflects whether a question can distinguish between passing and failing candidates (overall). We use 27% discrimination, which is a measure that 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

	Mean value						
	Jun 18 to Jan 20	**Jan 19 to Nov 20**	**Jun 19 to Mar 21**	**Jan 20 to Oct 21**	**Mar 22 to Oct 24**		**Long-run average (Jan 17 to Oct 24)**
Paper I discrimination	0.69	0.75	0.75	0.78	0.65		0.72
Paper II discrimination	0.87	0.87	0.89	0.89	0.64		0.83
Overall average discrimination	0.78	0.81	0.82	0.83	0.65		**0.78**

Our target for discrimination at paper level is >0.5

Chairs' summary 2021-24:

The available data and monitoring undertaken at each diet indicates that the exam continues to perform well psychometrically. We have consistently excellent reliability statistics that meet our targets and are in line with published best practice. Our SEM has been creeping up over the last series of exam 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, and this review notes that these have been maintained at a level considerably in excess of our target of 0.5.

Overall, these data are very reassuring, but we need to remain cognisant of our rising SEM and understand the reasons for the levels seen in recent exams. No immediate action is required, but this does need continued, close monitoring.