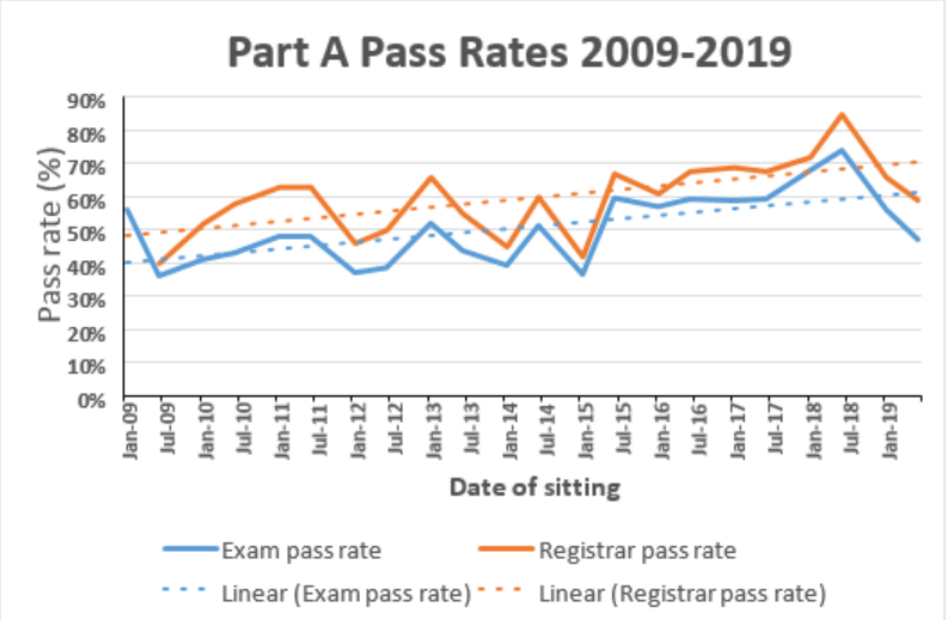
# Diplomate Exam monitoring and performance

**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.[1]
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). Assessment experts do not provide a guide to an 'ideal' SEM, but we have set a standard of <3.0.[2] [Note: SEM is not the same as the Standard error of the mean]



The figure on the previous page indicates a rising trend in pass rates with inter-diet variability. It is important to be aware that modified Angoff standard setting does not of itself remove pass rate variability. Instead, a number of factors including candidate variability will contribute to this. Furthermore, unlike large cohort undergraduate medical examinations, this exam is currently taken by approximately 70-80 candidates, which inevitably increases the variability observed above.

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** | | | | |
|---|---|---|---|---|---|
| | **Jan 2017 to Jun 2018** | **Jun 2017 to Jan '19** | **Jan 2018 to Jun 2019** | | **Long-run average (Jan '17 to Jun '19** |
| Paper I Cronbach alpha | 0.82 | 0.84 | 0.84 | | 0.84 |
| Paper I G-coefficient | 0.80 | 0.82 | 0.82 | | 0.82 |
| Paper II Cronbach alpha | 0.72 | 0.76 | 0.76 | | 0.73 |
| Paper II G coefficient | 0.71 | 0.74 | 0.74 | | 0.72 |
| **Exam reliability (Cronbach alpha)** | **0.87** | **0.89** | **0.89** | | **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** | | | | |
|---|---|---|---|---|---|
| | **Jan 2017 to Jun 2018** | **Jun 2017 to Jan '19** | **Jan 2018 to Jun 2019** | | **Long-run average (Jan '17 to Jun '19** |
| SEM Paper 1 | 3.23 | 3.25 | 3.43 | | 3.35 |
| SEM Paper 2 | 3.58 | 3.80 | 3.73 | | 3.69 |
| SEM Exam | **2.98** | **2.55** | **2.56** | | **2.87** |

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.

**Table 3: Examiner performance** – intra-class correlation (note calculated first in Jun 2017):

| | Four diet rolling average | | | | |
|---|---|---|---|---|---|
| | **Jan 2017 to Jun 2018** | **Jun 2017 to Jan '19** | **Jan 2018 to Jun 2019** | | **Long-run average (Jan '17 to Jun '19** |
| Average ICC across all examiner pairs | - | 0.83 | 0.84 | | 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.

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

| | Four diet rolling average | | | | |
|---|---|---|---|---|---|
| | **Jan 2017 to Jun 2018** | **Jun 2017 to Jan '19** | **Jan 2018 to Jun 2019** | | **Long-run average (Jan '17 to Jun '19** |
| Paper I mark correlations before agreement | 0.53 | 0.54 | 0.55 | | 0.56 |
| Paper I mark correlations after agreement | 0.75 | 0.76 | 0.77 | | 0.77 |
| Paper II mark correlations before agreement | 0.53 | 0.58 | 0.57 | | 0.57 |
| Paper II mark correlations after agreement | 0.74 | 0.77 | 0.76 | | 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.
- 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.

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

| | Four diet rolling average | | | | |
|---|---|---|---|---|---|
| | **Jan 2017 to Jun 2018** | **Jun 2017 to Jan '19** | **Jan 2018 to Jun 2019** | | **Long-run average (Jan '17 to Jun '19** |
| Paper I facility | 76% | 79% | 75% | | 74% |
| Paper IIA facility | 59% | 48% | 49% | | 51% |
| Paper IIB facility | 55% | 64% | 46% | | 56% |
| Overall facility of examination | 68% | 67% | 64% | | 65% |

No target set

**Table 6: Question discrimination averaged across all questions**

| | Four diet rolling average | | | | |
|---|---|---|---|---|---|
| | **Jan 2017 to Jun 2018** | **Jun 2017 to Jan '19** | **Jan 2018 to Jun 2019** | | **Long-run average (Jan '17 to Jun '19** |
| Paper I discrimination | 0.49 | 0.54 | 0.63 | | 0.57 |
| Paper II discrimination | 0.71 | 0.74 | 0.86 | | 0.76 |
| Overall average discrimination | 0.60 | 0.63 | 0.74 | | 0.66 |

Our target for discrimination at paper level is >0.5

**Chairs' summary 2017-19:**

The above monitoring indicates that the exam performs very well psychometrically. We have consistently excellent reliability statistics, a reasonably low (and consistent) SEM at exam level, examiners have consistently excellent intra-class correlation coefficients, and good correlation at question level before and after agreement. The questions themselves have also been observed generally to have very good discrimination.

**References:**


1) Tavakol M & Dennick R: Making sense of Cronbach's alpha. Int J Med Educ 2011 (2) 53-55

2) Tighe J, McManus I, Dewhurst N, Chris L, Mucklow J. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP (UK) examinations. BMC Med Educ 2010, 10:40