PLAB Part 1
Annual report, 2016-2017

This report details the performance of Part 1 candidates and test items, over the four diets delivered between September 2016 and June 2017.

The report is organised in four sections:

1) *Part 1 outcomes*; this section reviews the pass rate for each diet and shows the total scores that were achieved by candidates, including total scores by the number of the attempt at Part 1 (see page 2);
2) *Candidate performance by demographic characteristics*; this section explores pass rates and score distributions by candidates' gender, age, country of primary medical qualification, and nationality (see page 6);
3) *Part 1 and IELTS scores*; this section shows the relationship between Part 1 and IELTS component scores (see page 10);
4) *Item analysis*; this section presents the difficulty and discrimination values for items, organised by item skill, topic and domain classifications (see page 12).

The key findings outlined in this report include:

- In this period, Part 1 was taken by 3429 candidates. This represents a significant increase in the number of Part 1 candidates assessed, compared to the previous year (+42.5%; 1630 candidates);
- Of these, 2993 (77.9%) were taking Part 1 for the first time, and 511 (13.3%) for the second time;
- Overall, 70.2% of Part 1 candidates secured a pass. The pass rate was highest among those taking Part 1 for the first time, and decreased with every subsequent attempt;
- The relationship between pass rate and the number of the attempt at Part 1 is consistent with, and supports, the introduction of a new threshold of a maximum of four attempts at Part 1 (and Part 2), from September 2017;
- The pass rate was relatively stable over this period with between six and seven out of every ten PLAB candidates securing a pass for Part 1 during each diet;
- The Cronbach's alpha coefficient for every diet exceeded the minimum requirement for high stakes examinations, and was consistent across the diets;
- On average, female candidates gained marginally more passes and a marginally higher mean score than male candidates. This represents a marginal increase on the trend observed for 2015-2016;
- The majority of Part 1 candidates were aged between 27 and 38. Those aged 32 or under gained the highest pass rate, followed by those aged 52 and over (which is the same trend as observed in 2015-2016);
- Nearly half of all candidates (46%) identified their ethic origin as Asian or Asian British. The highest pass rate (82.4%) was secured by candidates identifying their ethnic origin as White, while the lowest pass rate (64.1%) was gained by those who did not specify their ethnic origin;
- As has been observed in previous years, there was a statistically significant, positive, but weak relationship between Part 1 scores and IELTS scores, whereby higher IELTS scores tend to be associated with higher Part 1 scores, but not exclusively so;
- The distribution of item difficulty and discrimination values is consistent with that observed in previous years. The item performance trends suggest that the items are performing as intended and are effective at discriminating between more and less able candidates.

SECTION 1

Part 1 outcomes

Table 1 shows that the pass rate has remained relatively consistent across diets, while the size of the cohort (and, therefore, it may be assumed, the spread of ability) has fluctuated, from a maximum of 1675 candidates in March, 2017, to 399 candidates in September, 2016.

Table 1. Part 1 results across diets (count and percent).

Diet[1]		Part 1 result		Total
		Fail	Pass	
Sept 16	**N**	137	262	399
	%	34.3	65.7	100.0
Nov 16	**N**	306	987	1293
	%	23.7	76.3	100.0
March 17	**N**	550	1125	1675
	%	32.8	67.2	100.0
June 17	**N**	150	320	470
	%	31.9	68.1	100.0
Total	**N**	1143	2694	3837
	%	29.8	70.2	100.0

The following table shows the descriptive statistics for the scores achieved by passing and failing candidates, and all candidates combined. It shows that:

- On average, passing candidates scored between 8-10% above the pass mark for each diet;
- Failing candidates, on average, scored approximately 8-9% below the pass mark;
- The mean scores achieved by all candidates were consistent over diets, with the exception of the November diet, which appeared one or two items easier – among the passing candidates only – than the other diets in this period;
- The 'Mean-pass' values represent the mean score *minus* the pass mark. In the case of the September and March diets, the mean score and pass mark were very close;
- The dispersion statistics (the standard deviation (SD), variance, range, and inter-quartile range (IQR)) suggest that candidates' total scores are quite tightly clustered in the mid-range, with very few candidates gaining a total score that is either very low or very high (as we would expect). The observed mark range was between 27-88%, with an average standard deviation of 10%;
- The Cronbach's alpha reliability coefficient for each diet is above the threshold typically considered to be necessary for high stakes examinations and is consistent with the coefficients produced in the previous year (an average of .91 for 2016-2017, compared with .92 previously).

[1] Percentage pass marks and total scores adjusted for suppressed items.

Table 2. Descriptive scores by diet (%).

	Sept			Nov			March			June		
Pass mark	63.45			63.27			63.64			62.80		
Candidates	**All**	**Pass**	**Fail**	**All**	**Pass**	**Fail**	**All**	**Pass**	**Fail**	**All**	**Pass**	**Fail**
Mean	66.20	71.75	55.59	68.83	73.16	54.86	66.28	71.77	55.05	66.30	72.27	53.57
Mean-pass	2.75	8.30	-7.86	5.56	9.89	-8.41	2.64	8.13	-8.59	3.22	9.47	-9.23
95% CI Lower	65.28	71.13	54.62	68.30	72.82	54.12	65.81	71.48	54.45	65.31	71.61	52.36
95% CI Upper	67.12	72.38	56.55	69.36	73.51	55.60	66.75	72.06	55.64	67.29	72.93	54.78
Median	67.51	71.57	56.85	70.41	73.47	56.63	67.68	71.72	56.57	67.84	71.86	55.53
Variance	87.35	26.25	32.32	95.51	31.09	43.24	95.03	24.96	50.53	118.62	36.04	56.62
SD	9.35	5.12	5.69	9.72	5.58	6.58	9.75	5.00	7.11	10.89	6.00	7.52
Min	29.95	63.45	29.95	31.12	63.27	31.12	26.77	63.64	26.77	27.64	62.81	27.64
Max	84.77	84.77	62.94	87.76	87.76	62.76	87.37	87.37	63.13	87.44	87.44	62.31
Range	54.82	21.32	32.99	56.64	24.49	31.64	60.60	23.73	36.36	59.80	24.63	34.67
IQR	13.70	8.12	8.38	12.24	8.67	9.18	12.62	8.08	9.60	14.07	9.54	9.18
Alpha	0.89			0.91			0.91			0.92		

The following chart shows the observed consistency in the distribution of scores by diet. As noted above, across the diets, even the most able candidates were unable to access more than 87% of items; while the least able candidates were able to access at least 27% of items. The November diet produced a marginally higher mean score, suggesting that this diet included some marginally more able candidates, or a small number of marginally more accessible items.

Chart 1. Distribution of scores (%) by diet.

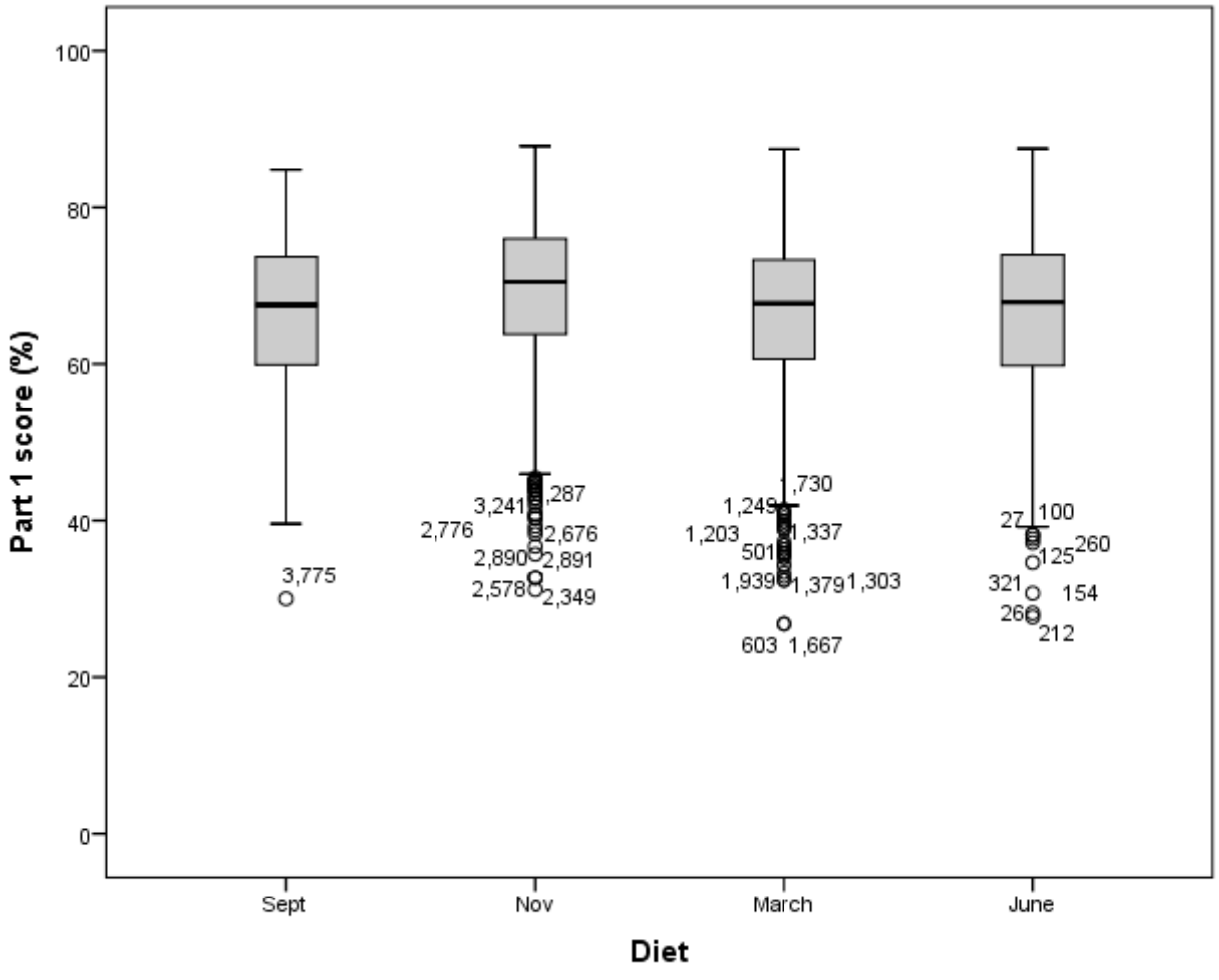
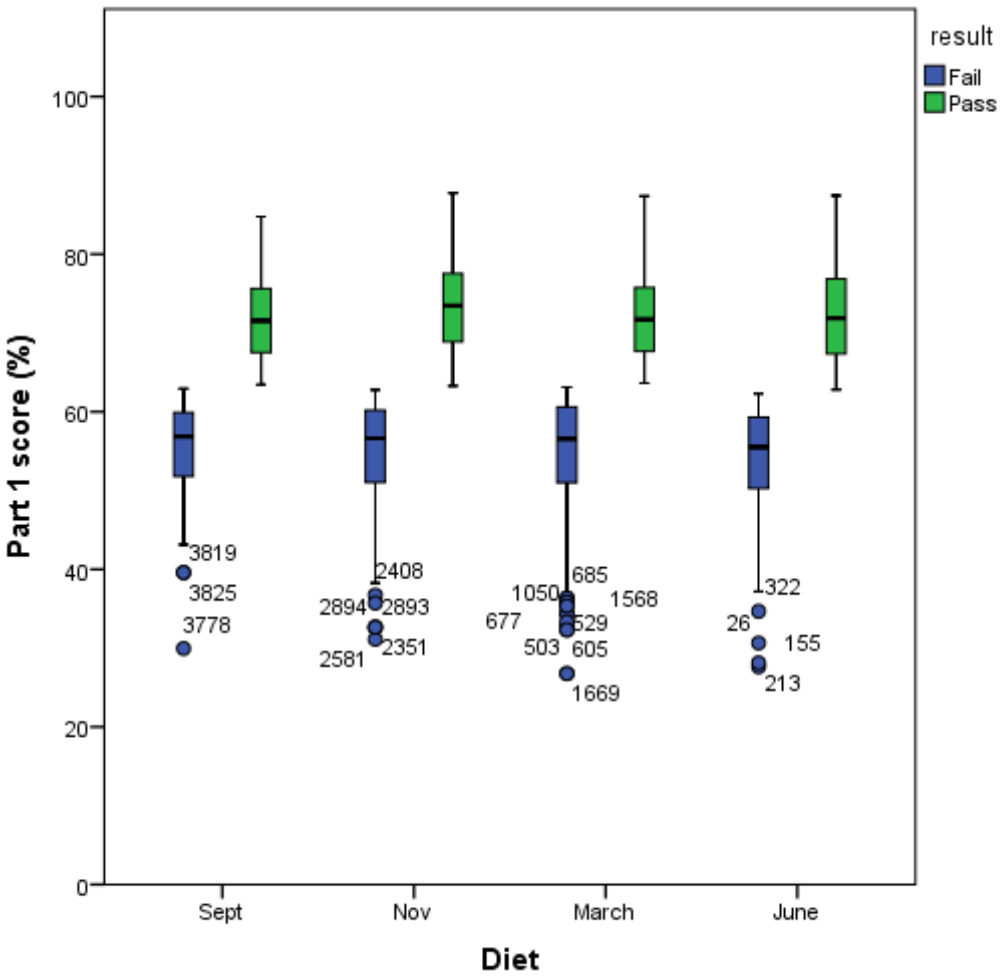


Chart 2 shows the same information but separated for passing and failing candidates. Again, the consistency is clear, with a similar spread of scores and mean score across both groups for each diet. The exception to this is the September diet, where the score distributions were more compressed than was observed for the other diets.

Chart 2. Distribution of scores (%) by diet and outcome.



The following table shows the mean score by the number of the attempt at Part 1. It shows that the likelihood of securing a pass is highest on the first take. Of those going through to a second, third and fourth take, the likelihood of securing a pass remains over 50%. However, likelihood of success falls quite significantly on the fifth attempt, and remains low for all subsequent attempts. The mean scores among the candidates on the fifth, or more, attempt, remain relatively stable, at approximately 53-59%, but with smaller standard deviations, suggesting more clustering of candidates below the pass mark.

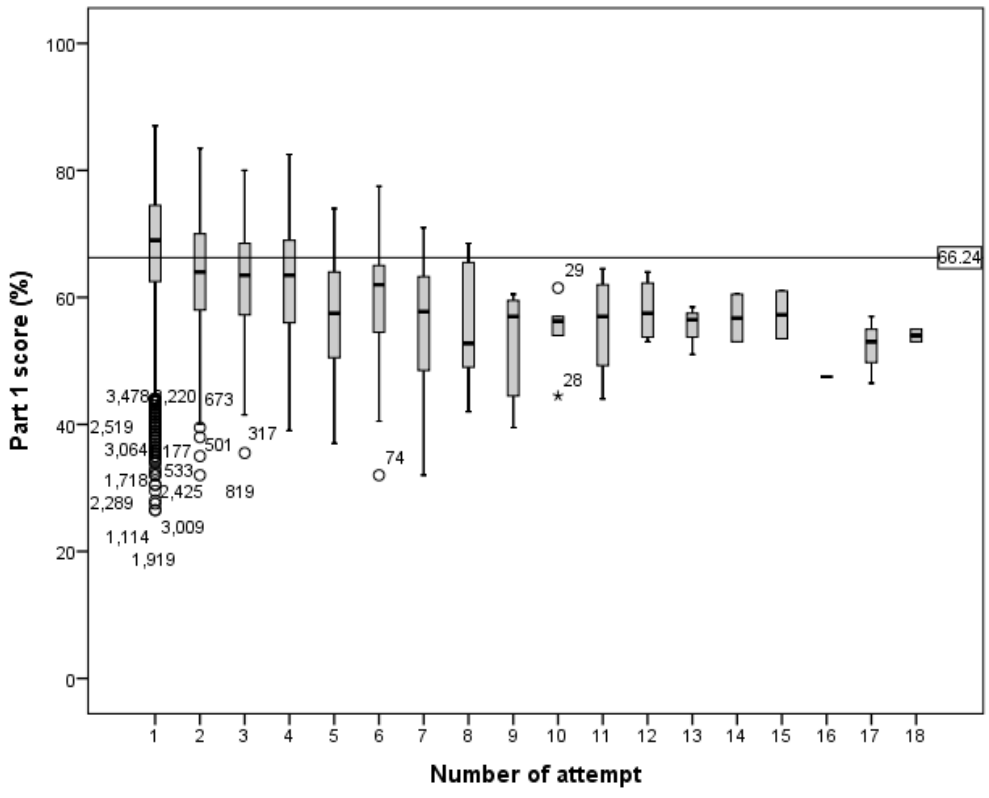
In September, 2017, a new threshold was set for the maximum number of attempts at Parts 1 and 2. This was set at four attempts, and is supported by the trends observed in 2015-2016, as shown below in Table 3 and Chart 3.

Table 3. Pass rate and Part 1 scores (%) by number of attempt (2016-2017 diets combined).

Attempt number	Candidates	Mean	SD	Min	Max	Pass (%)	Fail (%)
1	2993	67.4	9.6	26.5	87.0	74.9	25.1
2	511	63.4	8.8	32.0	83.5	58.7	41.3
3	152	62.5	8.4	35.5	80.0	57.9	42.1
4	53	62.0	9.9	39.0	82.5	54.7	45.3
5	37	57.1	9.3	37.0	74.0	32.4	67.6
6	29	58.8	9.8	32.0	77.5	44.8	55.2
7	16	54.7	11.2	32.0	71.0	25.0	75.0
8	10	55.6	9.0	42.0	68.5	30.0	70.0
9	6	53.0	8.8	39.5	60.5	0.0	100.0
10	6	54.9	5.7	44.5	61.5	0.0	100.0
11	7	55.4	8.2	44.0	64.5	14.3	85.7
12	4	58.0	5.1	53.0	64.0	25.0	75.0
13	3	55.3	3.9	51.0	58.5	0.0	100.0
14	2	56.8	5.3	53.0	60.5	0.0	100.0
15	2	57.3	5.3	53.5	61.0	0.0	100.0
16	1					0.0	100.0
17	3	52.2	5.3	46.5	57.0	0.0	100.0
18	2	54.0	1.4	53.0	55.0	0.0	100.0

Following on from Table 3, the chart below shows the spread of scores by the number of the attempt and in relation to the grand mean score (the mean score across all attempts at Part 1 in 2016-2017, which is dominated by first attempts). This shows that, among these cohorts, mean scores clearly decrease up to the fifth attempt and remain low thereafter. A Pearson correlation coefficient confirms that the relationship between these two variables is negative and statistically significant but fairly weak (r^2=-.23, p=.000).

Chart 3. Part 1 scores (%) by the number of the attempt.



SECTION 2

Part 1 performance by candidates' demographic characteristics

The next section explores Part 1 performance by the following selected demographic characteristics of trainees within each test cohort: gender, age, ethnic origin, country of primary medical qualification, and nationality. This section focuses on first take candidates only, and excludes from the analysis all re-take candidates. The total number of candidates per administration and by each characteristic is therefore determined by the proportion of first take candidates combined with the availability or completeness of demographic profiles for individual candidates.

Gender

The following table shows the distribution of pass and fail outcomes, and mean scores by candidates' gender (where specified). It shows that, across the 2016-2017 diets, female candidates gained a pass rate that was 3.3% higher than the male candidates, and a mean score that was 1.21% higher than males. Excluding the top and bottom 5% from each distribution, the difference reduces by only 0.14%, suggesting that the difference is not notably attributable to the performances of the outlier candidates. The differences observed between the genders for 2016-2017 represents a marginal increase on those observed in 2015-2016 (female pass rate +1.6% and mean score +0.7% for 2015-2016).

Table 4. Distribution of Part 1 pass and fail outcomes and total scores (%) by candidates' gender.

Part 1 result and score (%)		Male	Female
N		1780	1647
Part 1 result: Pass (%)		70.8	74.1
Part 1 result: Fail (%)		29.2	25.9
Part 1 score: Mean		67.93	69.14
95% Confidence Interval for Mean	**Lower Bound**	67.48	68.71
	Upper Bound	68.38	69.57
5% Trimmed Mean		68.56	69.63
Median		69.39	70.35
Variance		94.65	79.59
Std. Deviation		9.73	8.92
Minimum		26.77	28.14
Maximum		87.37	87.76
Range		60.60	59.62
Interquartile Range		11.36	10.73

An independent samples t-test was used to explore whether the observed difference in mean score was statistically significant. The results suggested that there was a systematic effect whereby, on average, female candidates gained marginally higher Part 1 scores than male candidates ($t_{3425} = -3.803$, p=.000).

Age

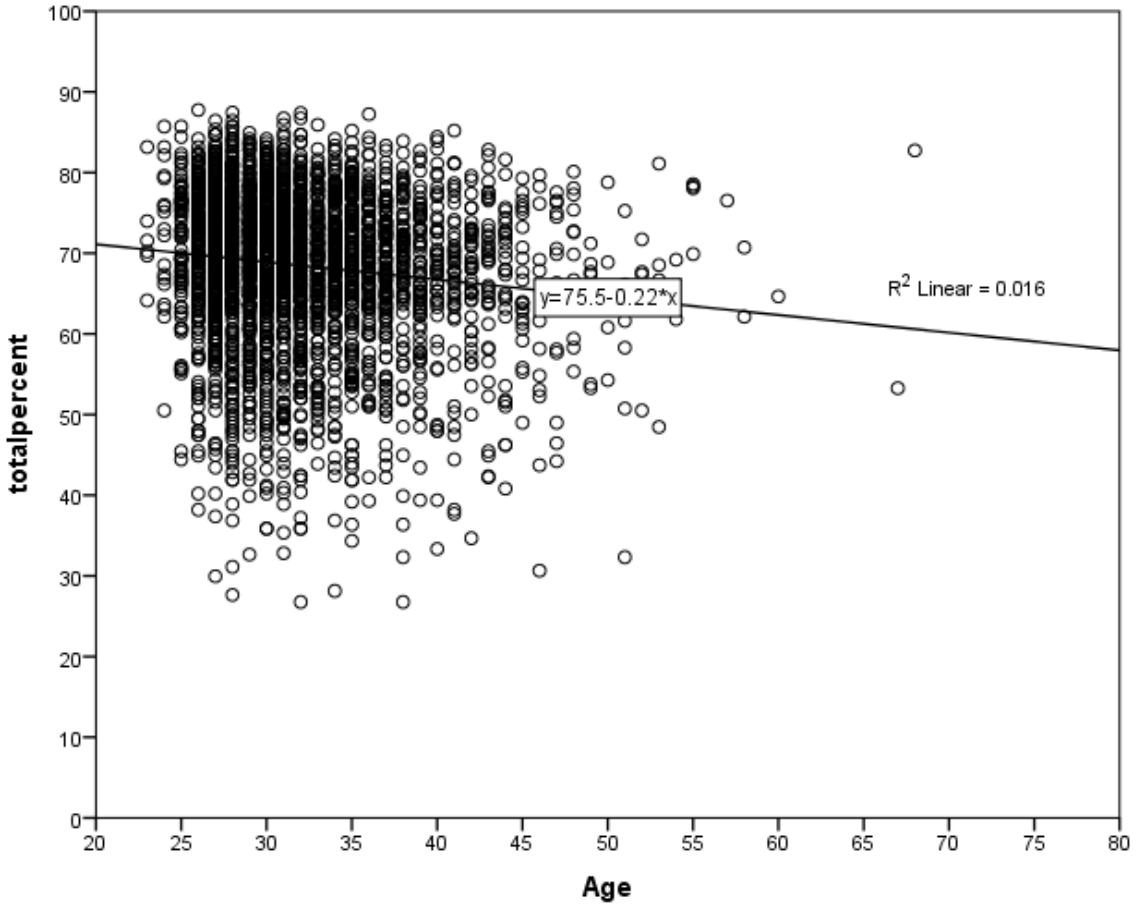
Table 5 shows how the candidates were distributed across six age bands, with most candidates aged between 27 and 38 (79.6%). The proportion of pass and fail outcomes, and the mean scores, indicate a general trend whereby performance, and likelihood of securing a pass, decreases with age. However, this holds true only up to age 52 and above, where performance increases once again (albeit based on a much smaller group of candidates – for 2016-2017, only 21 candidates were in this age band). The group under-performing relative to all others is aged 45-51, although this group is also fairly small.

Table 5. Distribution of Part 1 pass and fail outcomes and total scores (%) by candidates' age.

Age band	N	Part 1 result (%)		Part 1 score (%)					
		Pass	Fail	Mean	SD	Min	Max	Range	IQR
21-26	320	78.1	21.9	70.39	8.73	38.19	87.76	49.57	10.63
27-32	1910	74.4	25.6	69.07	9.19	26.77	87.44	60.67	10.62
33-38	818	69.9	30.1	67.61	9.34	26.77	87.24	60.47	10.66
39-44	275	64.7	35.3	66.32	10.12	33.33	85.20	51.87	10.84
45-51	83	53.0	47.0	64.65	10.47	30.65	80.10	49.45	14.94
52 >	21	71.4	28.6	68.24	9.50	48.47	82.74	34.27	13.91

A Pearson correlation between actual age and total Part 1 score (%) was negative and statistically significant, but weak, as we would expect given the trend among the candidates aged 52 and above, as noted in Table 5 (r^2 = -.125, p=.000). The relationship is shown in the scatterplot below.

Chart 4. Scatterplot of candidates' age and Part 1 score (%).



Ethnic origin

The following table shows the Part 1 outcomes and descriptive scores by candidates' ethnic origin. It shows that although the candidates specifying their ethnic origin as 'White' comprise one of the smallest groups, the likelihood of securing a pass, and their achieved mean score, exceed those of the other groups. The next highest performing group comprises 'Other ethnic groups' where the likelihood of obtaining a pass is -7.6% and the mean score is -1.76% that of the 'White' candidates. The most frequently identified ethnic group is 'Asian or Asian British' (46.0%). This group of

candidates secured a pass rate that is equal to the pass rate of all candidates combined (70.2%) and a mean score that exceeds the combined mean score across diets (68.02% compared with 66.9% across all candidates combined).

Table 6. Distribution of Part 1 pass and fail outcomes and total scores (%) by candidates' ethnic origin.

Ethnic origin	Candidates		Part 1 result (%)		Part 1 score (%)					
	N	%	Pass	Fail	Mean	SD	Min	Max	Range	IQR
Asian or Asian British	1576	46.0	70.2	29.8	68.02	9.68	26.77	87.76	60.99	11.36
Black or Black British	992	28.9	74.3	25.7	68.56	8.87	30.65	87.37	56.72	10.24
Mixed	48	1.4	66.7	33.3	68.27	9.51	45.45	84.77	39.32	15.18
Not stated	147	4.3	68.7	31.3	67.67	10.20	27.64	84.69	57.05	11.62
Other Ethnic Groups	412	12.0	75.2	24.8	69.64	9.23	26.77	87.44	60.67	11.11
Unspecified	78	2.3	64.1	35.9	67.07	9.23	35.86	84.18	48.32	9.82
White	174	5.1	82.8	17.2	71.40	8.15	32.83	85.71	52.88	10.47

Country of primary medical qualification (PMQ)

Table 7 shows the pass rate and mean scores achieved by the range of countries in which the Part 1 candidates gained their PMQ[2]. In the majority of cases, the numbers of candidates are very small and the pass rate and mean score, therefore, should be interpreted with some caution. As has been the case in previous years, the dominant countries in which Part 1 candidates qualified are (in order) Nigeria, Pakistan, India and Egypt. Among the groups achieving perfect, or very high, pass rates are Australia, Kenya, New Zealand, Oman, Turkey and Zambia. The highest mean score was achieved by candidates qualifying in Oman, followed by New Zealand, Sri Lanka and Brazil. The candidates gaining the lowest mean score qualified in Romania.

Table 7. Distribution of Part 1 pass and fail outcomes and mean scores (%) by candidates' country of primary medical qualification.

Country of PMQ	N	Part 1 result		Part 1 score (%)	
		Pass (%)	Fail (%)	Mean	SD
Afghanistan	6	50.0	50.0	68.17	10.97
Algeria	6	83.3	16.7	67.04	8.57
Armenia	6	66.7	33.3	69.65	5.79
Australia	40	90.0	10.0	72.67	6.95
Bahrain	31	77.4	22.6	67.35	9.09
Bangladesh	103	78.6	21.4	68.57	10.28
Belarus	8	87.5	12.5	72.87	8.14
Brazil	15	86.7	13.3	74.15	8.93
China	105	54.3	45.7	63.92	10.39
Egypt	217	71.9	28.1	68.33	9.16
Ghana	41	85.4	14.6	72.98	5.87
Grenada	14	78.6	21.4	67.42	6.77
Hungary	12	58.3	41.7	68.04	8.83
India	389	66.1	33.9	67.25	10.24
Iran, Islamic Republ	13	69.2	30.8	69.52	8.19
Iraq	41	75.6	24.4	72.14	8.31
Ireland	34	82.4	17.6	70.23	7.56
Jamaica	11	81.8	18.2	68.58	9.55

[2] Countries with fewer than five entries are excluded to ensure candidates are not identifiable by the data presented.

Country of PMQ	N	Part 1 result		Part 1 score (%)	
		Pass (%)	Fail (%)	Mean	SD
Jordan	19	78.9	21.1	71.88	8.37
Kenya	12	91.7	8.3	73.54	9.21
Libyan Arab Jamahiri	29	69.0	31.0	68.63	10.03
Malaysia	7	85.7	14.3	69.23	9.59
Mauritius	16	75.0	25.0	68.37	8.68
Myanmar	50	88.0	12.0	73.15	7.88
Nepal	32	75.0	25.0	69.05	9.50
New Zealand	7	100.0	0.0	75.63	5.96
Nigeria	845	74.6	25.4	68.55	8.66
Oman	8	100.0	0.0	76.29	5.41
Pakistan	758	73.0	27.0	68.41	9.13
Philippines	21	71.4	28.6	67.24	8.14
Poland	11	72.7	27.3	68.82	5.37
Romania	14	50.0	50.0	58.09	17.49
Russian Federation	48	64.6	35.4	66.21	9.46
Saint Kitts And Nevi	8	62.5	37.5	68.33	11.00
Saudi Arabia	12	58.3	41.7	60.71	12.98
South Africa	45	88.9	11.1	71.72	8.17
Sri Lanka	18	77.8	22.2	74.97	6.55
Sudan	76	55.3	44.7	66.93	10.77
Syrian Arab Republic	10	80.0	20.0	69.42	10.97
Turkey	18	100.0	0.0	73.56	5.20
Uganda	5	40.0	60.0	66.16	9.36
Ukraine	72	66.7	33.3	67.33	9.00
United Arab Emirates	48	66.7	33.3	69.02	10.28
Yemen	5	80.0	20.0	75.92	5.08
Zambia	5	100.0	0.0	67.54	2.90
Zimbabwe	23	82.6	17.4	70.19	7.79

Nationality

Table 8 shows the same information as above, but by candidates' nationality. The nationalities most frequently recorded in 2016-2017 were, in order, Nigerian, Pakistani, Indian and British. The nationalities gaining the most Part 1 passes were Turkish, Zambian, New Zealander, and Myanmarian. The nationalities gaining the fewest Part 1 passes were Afghan and Saudi Arabian.

Table 8. Distribution of Part 1 pass and fail outcomes and mean scores (%) by candidates' nationality.

Nationality	N	Part 1 result		Part 1 score (%)	
		Pass (%)	Fail (%)	Mean	SD
Afghan	5	40.0	60.0	64.56	13.31
American	35	68.6	31.4	69.40	8.70
Australian	29	82.8	17.2	70.68	7.09
Bahraini	17	76.5	23.5	67.07	9.50
Bangladeshi	77	80.5	19.5	69.64	10.05
Brazilian	14	85.7	14.3	74.67	9.03

Nationality	N	Part 1 result		Part 1 score (%)	
		Pass (%)	Fail (%)	Mean	SD
British	249	70.3	29.7	68.31	9.53
Burmese	5	60.0	40.0	68.76	16.21
Canadian	81	65.4	34.6	67.22	8.20
Chinese	12	58.3	41.7	69.40	7.19
Egyptian	189	74.6	25.4	68.82	8.68
Filipino	17	76.5	23.5	68.46	7.69
Ghanaian	28	75.0	25.0	70.46	8.08
Indian	439	64.7	35.3	66.75	10.37
Iranian	13	76.9	23.1	72.65	8.45
Iraqi	38	84.2	15.8	74.17	7.68
Irish	8	75.0	25.0	67.60	7.40
Jamaican	10	70.0	30.0	65.32	9.36
Jordanian	21	71.4	28.6	68.08	13.26
Kenyan	17	76.5	23.5	70.58	13.84
Libyan	27	74.1	25.9	69.16	9.88
Malaysian	20	80.0	20.0	71.23	10.16
Mauritian	25	76.0	24.0	67.88	10.46
Myanmar	46	91.3	8.7	73.61	6.47
Nepalese	46	76.1	23.9	68.19	10.88
New Zealander	12	91.7	8.3	75.57	5.25
Nigerian	896	74.6	25.4	68.69	8.59
Pakistani	695	72.4	27.6	68.12	9.38
Palestinian	8	87.5	12.5	71.84	9.17
Russian	7	100.0	0.0	70.73	5.03
Saudi Arabian	6	33.3	66.7	56.22	10.54
Singaporean	6	83.3	16.7	71.70	6.52
South African	44	77.3	22.7	69.59	10.22
Sri Lankan	50	62.0	38.0	68.58	9.51
Sudanese	42	59.5	40.5	66.08	11.01
Syrian	13	84.6	15.4	68.21	11.76
Trinidadian	13	53.8	46.2	65.19	6.77
Turkish	15	100.0	0.0	73.56	5.45
Ukrainian	8	50.0	50.0	69.73	7.79
Yemeni	5	60.0	40.0	75.49	8.72
Zambian	5	100.0	0.0	67.54	2.90
Zimbabwean	26	84.6	15.4	70.47	7.49

SECTION 3

Part 1 and IELTS scores

The table below shows Part 1 pass rate by candidates' achieved IELTS scores, across the individual IELTS components and the combined score. The table shows that, as we might expect, there is mostly good consistency between Part 1 pass rates and IELTS scores, with higher scores being associated with higher pass rates. The exception to this is the Speaking component which produces a curvilinear effect, with the lowest and highest Speaking scores being associated with the highest Part 1 pass rates, while the lowest Part 1 pass rates were achieved by those achieving the mid-range Speaking scores.

Table 9. Distribution of Part 1 pass and fail outcomes (%) by candidates' IELTS scores.

IELTS score	Reading			Writing		
	Fail	Pass	Total	Fail	Pass	Total
	%	%	N	%	%	N
6.0				100.0	0.0	1
6.5	100.0	0.0	1	33.3	66.7	3
7.0	36.3	63.7	570	29.4	70.6	2070
7.5	32.3	67.7	697	25.8	74.2	989
8.0	29.1	70.9	660	21.0	79.0	205
8.5	22.9	77.1	870	10.8	89.2	37
9.0	17.3	82.7	509	0.0	100.0	2
All	27.6	72.4	3307	27.6	72.4	3307

	Listening			Speaking			Combined		
	Fail	Pass	Total	Fail	Pass	Total	Fail	Pass	Total
	%	%	N	%	%	N	%	%	N
6.0									
6.5	100.0	0.0	1						
7.0	34.8	65.2	181	25.5	74.5	1049	100.0	0.0	1
7.5	33.8	66.2	538	25.6	74.4	1003	31.4	68.6	1393
8.0	28.0	72.0	804	30.2	69.8	703	25.7	74.3	1507
8.5	26.8	73.2	1119	34.7	65.3	349	21.7	78.3	387
9.0	21.2	78.8	664	26.6	73.4	203	10.5	89.5	19
All	27.6	72.4	3307	27.6	72.4	3307	27.6	72.4	3307

Following on from the table above, Table 10 shows the mean IELTS scores of those passing and failing Part 1. It shows that, again with the exception of the Speaking component, the candidates passing Part 1 gained higher IELTS scores than those failing Part 1. The 'Diff' column shows the magnitude of the difference in scores, which in all cases is small.

A correlation coefficient is also shown for each component and overall to show the strength and direction of the relationship between the Part 1 scores and IELTS scores. With the exception of the Speaking scores, all components produce a statistically significant relationship with Part 1 scores, but the strength of the linear relationships is weak, with the coefficients ranging only from .07 to .17. The correlation for the Speaking component is close to zero meaning that there is no linear relationship with the Part 1 scores - as we would expect given that speaking skills are not utilised nor assessed in Part 1.

Table 10. IELTS scores by Part 1 outcome.

	Part 1 Pass		Part 1 Fail		Diff (Fail-Pass)	R^2
	Mean	SD	Mean	SD		
Reading	8.07	0.67	7.85	0.65	-0.21	.17**
Writing	7.24	0.35	7.19	0.30	-0.05	.07**
Listening	8.27	0.57	8.15	0.58	-0.12	.11**
Speaking	7.63	0.59	7.69	0.60	0.06	-.02
Overall	7.87	0.35	7.81	0.33	-0.06	.11**

**Correlation is statistically significant at .01 level.

SECTION 4

Part 1 item analysis

This section shows the proportions and performances (item difficulty and discrimination) of all Part 1 items used in 2016-2017. The items are categorised by the following categories: skill, topic, and domain (see Appendix for item analysis by diet). The section ends by exploring the consistency of the observed item difficulty with the intended or perceived item difficulty, as indicated by the standard setting scores allocated to each item.

The following table shows that of the 800 items used in 2016-2017, the proportion of items across each skill was relatively consistent over the individual diets. The slight exception to this is the November diet, which included more diagnosis and genetics items, and fewer items relating to investigations and skills relating to long-term conditions. Across all diets, the most frequently assessed skill was diagnosis, which comprised a quarter of all items. The most challenging items (i.e. having the lowest difficulty values) assessed skills in epidemiology and physiology, while the most accessible items assessed skills in perioperative and pharmacology – however, these all represented small item groups.

Table 11. Item skill: distribution of items by diet, and facility and discrimination values.

Item skill	Distribution of items per diet (%)				2016-2017 diets combined				
	Sep-16	Nov-16	Mar-17	Jun-17	N	Facility	SD	Discrim	SD
Acute including emergency	11.5	11.0	10.0	17.0	99	.61	.24	.23	.13
Anatomy	4.5	7.0	4.5	4.0	40	.60	.18	.22	.09
Diagnosis	24.5	34.5	23.0	18.0	200	.69	.23	.22	.13
Epidemiology and health promotion	1.0	1.5	2.5	1.0	12	.54	.28	.17	.12
Genetics	1.5	3.0	1.0	1.0	13	.65	.32	.27	.16
Investigation: appropriate tests	11.0	7.0	13.5	13.0	89	.68	.21	.23	.14
Investigation: interpret results	10.0	9.0	10.5	8.5	76	.65	.26	.26	.12
Long-term	12.5	6.0	13.0	11.5	86	.69	.22	.26	.13
Pathology and microbiology	2.5	4.5	4.0	6.0	34	.63	.21	.27	.13
Perioperative	4.0	3.0	5.0	5.5	35	.73	.21	.25	.11
Pharmacology and biochemistry	4.0	4.0	3.0	3.0	28	.73	.17	.31	.08
Physiology	3.0	2.5	1.5	2.0	18	.55	.28	.23	.19
Social science and statistics	2.0	.5	1.0	1.0	9	.59	.33	.16	.11
Symptom relief and end of life	6.0	4.5	6.0	6.0	45	.69	.19	.31	.12
Unspecified	2.0	2.0	1.5	2.5	16	.68	.19	.13	.12
Total/average	100.0	100.0	100.0	100.0	800	.66	.23	.24	.13

Table 12 shows the same information as above, but by item topic. With a larger number of classifications, the numbers in each topic area tend to be small, meaning that the difficulty and discrimination values should be interpreted with some caution as they may conceal trends and variability at an individual item level. As in 2015-2016, the most frequently assessed topic was mental health, followed by digestive. By a small margin, and based on a small number of items, the most accessible items appeared to assess older adults, followed by a larger group of items focused on endocrine. On average, the most challenging items assessed the topics of respiratory and musculoskeletal. With a couple of exceptions, the average discrimination value for each topic area suggests that the items were effective in discriminating between the more and less able candidates. The exceptions are the ethical and professional, urological and child health items, where the relationship between total score and item score was not as strong.

Table 12. Item topic: distribution of items by diet, and facility and discrimination values.

Item topic	Distribution of items per diet (%)				2016-2017 diets combined				
	Sep-16	Nov-16	Mar-17	Jun-17	Total number	Facility	SD	Discrim	SD
Blood and lymph	3.0	2.5	2.5	2.0	20	.66	.25	.31	.09
Breast	1.5	1.5	1.0	2.0	12	.69	.28	.27	.11
Cardiovascular	7.5	5.5	7.0	8.0	56	.61	.24	.24	.12
Child health	4.0	3.0	2.5	1.5	22	.63	.29	.18	.13
Digestive	8.5	9.0	9.5	9.5	73	.70	.18	.21	.11
Endocrine	4.5	5.0	5.0	4.5	38	.77	.17	.30	.07
ENT	3.5	5.5	4.0	4.0	34	.69	.23	.26	.15
Ethical and professional	2.0	2.0	1.5	2.5	16	.65	.20	.11	.11
Eye	2.0	2.5	2.0	2.0	17	.62	.26	.24	.14
Genitourinary	2.5	3.0	2.0	2.0	19	.68	.20	.27	.13
Homeostatic	3.0	3.0	2.0	2.5	21	.74	.13	.29	.11
Infectious disease	3.5	3.5	4.5	4.0	31	.68	.20	.27	.14
Mental health	11.5	11.0	13.0	13.0	97	.71	.21	.25	.13
Musculoskeletal	5.5	5.0	4.5	4.5	39	.58	.22	.21	.12
Neurological	7.0	7.0	7.0	6.5	55	.64	.23	.28	.11
Older adults	1.0	.5	1.0	1.5	8	.78	.21	.25	.05
Pharmacological	1.0	1.0	1.0	.5	7	.76	.17	.29	.09
Renal	3.0	3.5	3.5	4.0	28	.67	.29	.24	.13
Reproductive	7.5	8.5	7.5	7.5	62	.68	.25	.24	.16
Respiratory	5.5	7.0	7.5	6.5	53	.57	.24	.23	.14
Seriously ill patient	7.5	6.5	7.0	5.5	53	.58	.28	.20	.15
Skin	2.0	2.5	2.0	2.5	18	.73	.19	.27	.09
Urological	3.0	1.5	2.5	3.5	21	.70	.20	.15	.16
Total/average	100.0	100.0	100.0	100.0	800	.67	.22	.24	.12

Of the three overarching domains assessed by Part 1, the majority of items concerned ‘Good clinical care: assessment’ (49.5%). The items in this domain appeared marginally more accessible than those in the other two domains, but across such a large number of items, there would undoubtedly be fluctuations in difficulty, with the easier items to some extent cancelling out the effects of the more challenging items. The standard deviations are similar across the three domains, suggesting that the range of difficulty is consistent.

The discrimination values were consistent across the domains. Again, allowing for the inevitable fluctuation in discrimination across individual items, it appears that, on the whole, the items were effective in discriminating between the more and less able candidates.

Table 13. Item domain: distribution of items by diet, and average facility and discrimination values.

Item domain	Distribution of items per diet (%)				2016-2017 diets combined				
	Sep-16	Nov-16	Mar-17	Jun-17	Total number	Facility	SD	Discrim	SD
Applying knowledge and experience to clinical practice	19.5	24.0	16.5	20.5	161	.65	.22	.24	.14
Good clinical care: assessment	46.5	51.5	46.5	39.0	367	.68	.23	.24	.13
Good clinical care: management	34.0	24.5	37.0	40.5	272	.66	.24	.25	.14
Total	100.0	100.0	100.0	100.0	800	.65	.22	.24	.14

Table 14 shows the distribution of items by patient age, with average difficulty and discrimination values. The majority of items related to adult patients (67.4%). With the exception of items where the patient's age was not relevant or specified, there are minimal differences between the categories in relation to item difficulty and discrimination. However, for the adult items particularly, the large number of items means that the averages reflect some degree of regression to the mean, with extreme high and low values cancelling each other out.

Table 14. Items classified by patient's age: distribution of items and average facility and discrimination values.

	Count	Percent	Facility	SD	Discrim	SD
Adult	539	67.4	.68	.21	.25	.12
Child	129	16.1	.65	.23	.21	.15
Elderly	98	12.3	.64	.26	.23	.13
Unspecified	34	4.3	.52	.22	.27	.12

All Part 1 items are subject to a modified Angoff process, whereby a number of trained raters negotiate and agree the level of perceived difficulty for each item, which is then combined with all other items and used to set the pass mark for the test as a whole. It is useful to conduct a comparison of the raters' perceptions of item difficulty against the candidates' observed experiences of item difficulty, to understand, in broad terms, the reliability of the rater judgements. It is essential to note, however, that the raters make their judgement of difficulty in relation to borderline candidates only, rather than the whole candidature. In including all candidates, the analysis below inevitably produces more differences than if 'borderline' candidates only had been included. Even so, taken as a whole, there is sound agreement between the perceived and observed difficulty values for each item, with a statistically significant correlation coefficient across all items where the Angoff score was available ($r^2 = .32$, $p = .000$, $n=520$).

Table 15 is a crosstabulation of perceived difficulty (Angoff score) by observed difficulty (facility value). It shows, for each Angoff score in the scale, whether the raters' perceptions of difficulty were above, below, or equal to the observed difficulty. As a general comment, the table shows that although the items covered nearly the full breadth of difficulty (i.e. on some items, fewer than 40% of candidates answered correctly, while, for other items, over 90% answered correctly), the raters rarely identified items as being very difficult or very easy. This may be a product of regression to the mean across the group of raters, a perception that borderline candidates will not find any of the items very easy or very difficult, or a reluctance to use the extremes of the Angoff scale.

The emboldened cells on the diagonal show the number and proportion of items where the raters' judgement of difficulty matched the candidates' experiences of item difficulty (across the whole ability spectrum, and not solely borderline candidates). An exact match was achieved for 16.5% of all items. Red-shaded cells indicate the items that were experienced as being more difficult than was expected by the raters (34.4% of all items). The green-shaded cells show the items that were found to be easier than the raters perceived them to be (47.9%). The number of items across the green-shaded cells is no doubt partly explained by the fact that, with a pass rate of 70.2%, the average candidate was performing above 'borderline' and was therefore more able than the candidate kept in mind for the Angoff process.

Table 15. Crosstabulation of perceived item difficulty and observed item difficulty.

Facility (observed difficulty)		Angoff Score (perceived difficulty)						Total
		4.0	5.0	6.0	7.0	8.0	9.0	
0-.39	N	1	11	51	15			78
	%	1.3	14.1	65.4	19.2			100.0
.40-.49	N		3	29	4			36
	%		8.3	80.6	11.1			100.0
.50-.59	N		**2**	25	14	3		44
	%		**4.5**	56.8	31.8	6.8		100.0
.60-.69	N		13	**41**	22	1		77
	%		16.9	**53.2**	28.6	1.3		100.0
.70-.79	N		14	52	**36**	5		107
	%		13.1	48.6	**33.6**	4.7		100.0
.80-.89	N		4	41	59	**7**	1	112
	%		3.6	36.6	52.7	**6.3**	.9	100.0
.90-1.00	N		3	19	33	11		66
	%		4.5	28.8	50.0	16.7		100.0
Total	N	1	50	258	183	27	1	520
	%	.2	9.6	49.6	35.2	5.2	.2	100.0

CONCLUDING COMMENTS

This report details the performance of items, tests and candidates across the four diets delivered in 2016-2017. The average pass rate of 70.2% is marginally down on that recorded for 2015-2016 (72.3%). However, the number of candidates has increased by 42.5% (a total of 1630 candidates). As such, the pass rate appears remarkably stable in the light of such an increase in the size of the candidature (on the whole – it was noted that the November 2016 diet produced a higher pass rate than the other diets in this period).

There are a number of trends that follow those observed in previous years. These include:

- The greater likelihood of candidates achieving a pass on their first take of Part 1, and the pass rate reducing with every subsequent attempt thereafter;
- A marginal difference in the attainment of males and females, with female candidates gaining slightly more pass outcomes and a slightly higher mean score than male candidates;
- Younger (aged 32 or below) and more mature (aged 52 and over) candidates out-performing other age groups;
- The rank ordering of performance trends by country of primary medical qualification and candidates' nationality includes many of the same countries of previous years, with, for example, the highest proportion of pass outcomes and highest mean scores being achieved by candidates from, or qualifying in, Australia, Myanmar, New Zealand and Oman;
- The performances of candidates specifying their ethnic origin as White is slightly different in 2016-2017 to that observed in 2015-2016 when White candidates were out-performed by other ethnic origin groups. This increase has occurred in the context of a much larger cohort, but reduced proportion of White candidates (from 6.6% in 2015-2016 to 5.1% in 2016-2017).
- The general trend that higher achieved IELTS scores is associated with higher Part 1 scores;
- The performance of items across diets, topic areas, skills, and domains indicates that items are performing as intended, producing facility and discrimination values that are consistently within an appropriate range, and producing score distributions and outcomes that are similar to those in 2015-2016;

All the performance indicators suggest that Part 1 continues to produce outcomes that are consistent with trends of previous years. Given the significant increase in the number of candidates noted for 2016-2017, it will be important to continue to monitor entries and any impacts on performance trends in future diets.

APPENDIX: ITEM FACILITY AND DISCRIMINATION BY DIET

Table 1. Item skill

Item skill	Sept-16				Nov-16				Mar-17				June-17			
	Facility	SD	Disc	SD	Facility	SD	Disc	SD	Facility	SD	Disc	SD	Facility	SD	Disc	SD
Unspecified	.54	.18	.13	.15	.56	.21	.06	.08	.79	.13	.10	.06	.82	.11	.18	.16
Acute including emergency	.68	.22	.21	.14	.55	.30	.19	.17	.56	.28	.21	.14	.58	.22	.25	.12
Anatomy	.54	.20	.22	.11	.61	.25	.27	.08	.66	.16	.23	.07	.61	.15	.22	.11
Diagnosis	.71	.20	.22	.13	.72	.23	.24	.13	.67	.25	.22	.13	.65	.24	.27	.12
Epidemiology and health promotion	.51	.04	.20	.08	.80	.03	.33	.12	.64	.30	.19	.13				
Genetics	.44	.40	.18	.20	.66	.30	.29	.14	.83	.17	.34	.16	.81	.04	.35	.06
Investigation: appropriate tests	.72	.18	.19	.15	.66	.21	.26	.11	.63	.25	.22	.15	.68	.24	.27	.11
Investigation: interpretation of results	.65	.24	.25	.13	.68	.27	.25	.15	.65	.25	.28	.10				
Long-term	.66	.20	.23	.15	.66	.16	.25	.11	.71	.25	.27	.13	.70	.18	.30	.09
Pathology and microbiology	.63	.22	.20	.09	.48	.25	.18	.13	.67	.24	.27	.15	.59	.20	.33	.13
Perioperative	.80	.13	.31	.13	.77	.14	.28	.21	.72	.24	.22	.09	.68	.24	.24	.11
Pharmacology and biochemistry	.75	.14	.32	.07	.83	.12	.26	.16	.81	.07	.31	.09	.51	.26	.28	.11
Physiology	.55	.31	.11	.17	.73	.29	.28	.13	.62	.30	.36	.04	.44	.23	.41	.11
Social science and statistics	.45	.36	.10	.09	.56	.21	.06	.08	.90	.03						
Symptom relief and end of life	.59	.21	.28	.12	.55	.30	.19	.17	.74	.12	.34	.11	.76	.18	.30	.13

Table 2. Item topic

Item topic	Sept-16				Nov-16				Mar-17				June-17			
	Facility	SD	Disc	SD	Facility	SD	Disc	SD	Facility	SD	Disc	SD	Facility	SD	Disc	SD
Blood and lymph	.65	.22	.30	.12	.74	.31	.32	.09	.72	.25	.31	.07	.46	.15	.34	.08
Breast	.56	.22	.26	.01	.72	.37	.23	.13	.91	.02	.36	.11	.61	.44	.27	.23
Cardiovascular	.63	.27	.23	.15	.55	.24	.22	.11	.64	.24	.25	.11	.60	.22	.24	.12
Child health	.49	.30	.18	.18	.67	.34	.15	.10	.76	.17	.19	.10				
Digestive	.65	.19	.20	.10	.70	.13	.22	.13	.75	.17	.23	.11	.66	.28	.14	.10
Endocrine	.75	.13	.29	.06	.83	.14	.32	.09	.79	.14	.30	.07	.71	.27	.30	.06
ENT	.79	.10	.31	.08	.65	.25	.23	.18	.58	.29	.24	.16	.78	.20	.30	.12
Ethical and professional	.54	.18	.13	.15	.56	.21	.06	.08	.79	.13	.10	.06	.82	.11	.18	.16
Eye	.81	.04	.22	.08	.60	.31	.29	.17	.58	.30	.20	.16	.46	.28	.22	.20
Genitourinary	.68	.26	.24	.18	.67	.16	.29	.10	.78	.16	.29	.09	.58	.29	.25	.21
Homeostatic	.75	.11	.34	.06	.76	.17	.22	.16	.69	.15	.28	.03	.72	.04	.34	.05
Infectious disease	.77	.11	.28	.10	.60	.24	.28	.15	.67	.25	.22	.18	.69	.16	.30	.11
Mental health	.67	.24	.21	.11	.81	.12	.29	.10	.67	.24	.23	.16	.71	.19	.30	.10
Musculoskeletal	.60	.18	.21	.08	.53	.27	.18	.17	.59	.23	.22	.11	.62	.19	.25	.10
Neurological	.63	.23	.23	.12	.66	.21	.30	.09	.66	.29	.28	.12	.63	.20	.31	.07
Older adults	.72	.35	.22	.05	.87	.05	.32	.06	.92	.05	.29	.06	.72	.25	.27	.03
Pharmacological	.80	.18	.23	.13	.72	.33	.26	.17	.78	.09	.31	.13				
Renal	.65	.33	.25	.10	.70	.26	.24	.15	.70	.27	.24	.14	.62	.28	.20	.13
Reproductive	.72	.18	.18	.16	.57	.29	.23	.14	.64	.29	.27	.15	.57	.29	.31	.21
Respiratory	.57	.22	.21	.17	.49	.29	.18	.16	.59	.25	.26	.15	.53	.22	.24	.12
Seriously ill patient	.65	.25	.16	.19	.78	.12	.22	.08	.60	.28	.24	.11	.60	.34	.24	.13
Skin	.81	.05	.34	.08	.89	.08	.24	.17	.62	.33	.22	.13	.70	.17	.30	.03
Urological	.66	.26	.09	.20	.74	.31	.32	.09	.63	.14	.13	.10	.69	.21	.22	.14

Table 3. Item domain

Item domain	Sept-16				Nov-16				Mar-17				June-17			
	Facility	SD	Disc	SD	Facility	SD	Disc	SD	Facility	SD	Disc	SD	Facility	SD	Disc	SD
Applying knowledge and experience to clinical practice	0.56	0.24	0.19	0.13	0.68	0.21	0.25	0.14	0.73	0.60	0.19	0.21	0.23	0.26	0.13	0.11
Good clinical care: assessment	0.70	0.21	0.22	0.13	0.71	0.22	0.25	0.12	0.25	0.27	0.12	0.14	0.68	0.66	0.24	0.22
Good clinical care: management	0.67	0.20	0.24	0.14	0.61	0.29	0.22	0.15	0.65	0.65	0.25	0.25	0.26	0.28	0.13	0.11