Review and advice regarding the ACEM Fellowship OSCE results from 2016.2

Professor Lambert Schuwirth

Prideaux Centre for Health Professions Education

Flinders University

September 2017

Part 1 Short background and brief

The 2016.2 OSCE results showed a considerable difference between the pass rates of candidates identified as Caucasian (CC) and those identified as Non-Caucasian Candidates (NCC). The pass rate for the CC cohort was rumoured to be 88% and that of the NCC was 6.8% (in fact they were 70.5 versus 13.5). Logically, this raises the question if this is the result of a form of assessment bias or whether there are other explanations possible. The brief for this review was therefore, to investigate the likelihood of either the existence of a bias or another explanation from a psychometrical/statistical point of view.

More concretely the requirements in the brief were:

- To run a reliability analysis of the examination and evaluate the score distribution in order to establish the number of bare failing and passing candidates of both candidate groups (CC and NCC).
- To perform a Chi-squared test on the pass/fail rates to estimate the likelihood of the difference in pass-fail percentages having occurred by chance.
- To establish whether there is statistical evidence against NCC candidates.
- To establish whether there is an examiner or station propensity to mark NCC candidates harder than CC candidates.
- To establish whether there are particular domains that have a propensity to be harder for NCC candidates than for CC candidates.
-

Additional queries were:

- Whether the College's current practice regarding the forms/rubrics/scoring are:
    - Standard practice across other postgraduate high-stakes examinations
    - Commensurate with the use of the borderline regression method
- Whether it is possible to review and remark the 2016.2 results or statistically correct any bias or stations that have contributed to the disparate outcomes
- To advise on what other options may be available.

Part 2 Considerations

2.1 Regarding the nature of the problem

The concern raised by some candidates that the results difference may be due to a form of bias in the examination is not unreasonable and the College's response to this concern shows that it takes the matter seriously. However, it will be extremely difficult to provide a clear and unambiguous answer to the College's queries. It is more or less like one equation with three unknowns. When one group of candidates performs markedly different from other groups on such examination there are typically three possible explanations:

1) the groups *are* markedly different with respect to the ability the examination purports to asses; so the examination is valid and the difference that it picked up is a true difference, or

2) the groups are not really different in ability and the examination is biased and therefore not valid, or

3) the finding is a one-off effect which is due to measurement error.

Unfortunately, it is not possible to disentangle these explanations with absolute certainty with the avialable data. Therefore, I have conducted a series of analyses to evaluate the likelihood of these three explanations from various perspectives.

An additional complexity in this problem is the difficulty when trying to demonstrate the absence of an effect. When analyses indicate that there *is* a bias the conclusion can be straightforward, but when no indication is found for any form of bias, it is important to also demonstrate that the analyses were sensitive enough to detect the bias if it had been present. In this review this issue is slightly easier because a bias will have to be identified that is big enough to have caused the considerable difference in pass rates and so it would have to emerge quite clearly from the data..

2.2 Regarding the treatment of the data

Intuitively one would be inclined to use normal inferential statistics to determine the significance of the difference between both groups. But in this case we do not want to infer whether the findings in these cohorts on this examination would generalise to a larger population of candidates or stations, so the standard inferences are not pertinent to this review. Therefore, where I have applied standard (parametric) statistical tests they are merely used to roughly gauge the likelihood of the effects being chance finding. For some queries I have combined this with a standard psychometric approach – which takes into account the specific measurement error component in the assessment – and more mainstream statistics – which are based more on the variance of the results-.

At this point it is good to repeat that the question posed is NOT whether ***any*** form of examination would be biased against ***any*** group of NCC candidates, but whether ***this*** examination was biased against ***this*** group of NCC candidates. Therefore in this report I am not attempting to make any inference about the 'population' of candidates with the parametric statistical analyses. When using psychometric analyses though there is always an automatic inference as to the true score or true score variance, i.e. there is a universe generalisation assumption. Yet, in both cases it is not done to evaluate the accuracy/reproducibility of the results but rather to gauge the likelihood of the effects having occurred by chance.

The volume of the data is not huge; it concerns 204 candidates each of whom 'sat' and OSCE of 15 stations. The candidates were subdivided into two cohorts. Cohort 1 consisted of 106 candidates and were presented with stations 1 to 15, cohort 2 consisted of 98 candidates who 'sat' stations 16 – 30. For this reason I have made the following decision for the analyses:

1) To use the simplest statistical/psychometric procedures requiring the lowest level assumptions. For the psychometric analyses I have used classical test theory only. For the purpose of this analysis Generalisability theory and Item response theory perspectives would not have been feasible or useful. Standard statistical analyses are kept simple (T-tests, Chi-square and descriptive statistics) as well.

2) To conduct all the analysis separately for both cohorts. It would be difficult to pool the data; not only were the stations different by content, also the number of double-marked stations and the division of subdomains differed. This way, the results of each cohort serves as a cross validation for the other.

2.3 Regarding the philosophy of OSCEs

OSCEs were developed in the mid-1970s by Ronald Harden and co-workers. In that era the dominant notion of assessment was one of 'measurement' of competence. The prevailing theory on competence held that it was best assessed by focusing on separate traits (typically: 'knowledge', 'skills', 'problem solving ability' and 'attitude') and that each of these traits could be measured generically and separately from each other (so, one could have skills without knowledge, and vice versa). The second assumption was that these traits were relatively stable, i.e. they were assumed not to change during the measurement. From this assumption of stability and generic nature of traits some design principles for assessment followed. One principle is that differential performance on assignments (in the case of an OSCE: the stations) of a candidate was most likely due to measurement error. To use an analogy: if from a vile of homogenised blood three subsamples are taken for a haemoglobin measurement, all three measurements should lead to the same value and if there are differences they are assumed to be due to measurement error. For OSCEs this has always been somewhat counter-intuitive, as it is not really easy to explain why a candidate who fails one skill should be allowed to compensate for this with good performance on another skill (in real practice a good knee examination does not make up for a bad abdominal examination).

Since then, our knowledge about the nature of medical competence and how to best assess it (and even the psychometric models) have changed dramatically and most of these changes are now influencing the way medical schools organise their assessment. However, there is still a widespread practice of traditional OSCEs both in medical schools and licensing and credentialing bodies.

This not necessarily bad practice. The context of licensing and credentialing is particularly high stakes and litigious. Stakeholder perception of correct and defensible practices may not always align with best evidence-based practice. I am highlighting this because in this report I am fully aware that although I could make suggestions which are based on best evidence from the literature, they might be politically, legally and PR-wise not yet sufficiently defensible in the context of licensing and fellowship examinations. However, if the College were to consider this it would require a long-term project and a carefully laid-out strategy. This, however, is beyond the scope of this report.

Part 3 Documentation provided

For this review I have received to following documentation:

- 2 data files (Excel). One with the results broken down by station (station scores) for both cohorts and one with the results broken down by curriculum domain.
- Individual descriptions of the 30 stations, with the candidate information, the examiner information, the role player information and in most cases an example of the score (rubric) form.
- PowerPoint slides with the examiner briefing.
- A copy of the examiner briefing form.
- A report written by Prof. Farmer about the 2016.2 examination.
- The Fellowship 2016.2 OSCE examination analysis report.

Part 4 Analyses

From the excel files I have created four different SPSS files for further analyses:

- Cohort 1 station scores.sav (containing the station scores and the origin of candidates and each of the examiners of cohort 1)
- Cohort 2 station scores.sav (containing the station scores and the origin of candidates and each of the examiners of cohort 2)
- Domains_cohort1 (containing the domain score and the origin of candidates of cohort 1)
- Domains_cohort2 (containing the domain score and the origin of candidates of cohort 2)

The tables with the data conversions are in appendix 1.

The following analyses were performed:

1 Reliability analysis of both cohorts and calculation of standard errors of measurement (SEM) and 95% confidence intervals (95% CI).
2 Reliabilities of the NCC and CC group separately for both cohort and the calculation of all SEMs and 95%Cis for both cohorts and both groups of candidates. Plus item analysis and comparison between NCC and CC candidates to determine whether there are specific stations with high differences.
3 Chi square test of candidate groups (NCC or CC) against passing or failing. For this analysis the cohort have been pooled.
4 T-test between the two candidate groups (NCC and CC) for descriptive purpose only and compare with overlap of 95% confidence intervals.
5 Determination of numbers of possible false-positive and false-negative results (score within a 95% CI around the cut-off score) and true-positive and true-negative results (scores outside the 95%CI around the cut-off score). Comparison between NCC and CC in both cohorts.
6 Calculation of the curriculum domain scores for both cohorts and comparison between the NCC and CC groups. For this I have used T-tests, not with the intent to make population

inferences but to scan for specific domains that would be more likely to produce a difference than others or whether the difference can be found across all domains (restricted to those domains that were examined with more than one station).

Given the number of analyses and the fact that the same question is often addressed using various analyses the process may appear to be a proverbial ‘fishing expedition’. But, like a physician who wants to demonstrate the absence of a disease uses the most sensitive armamentarium of diagnostic tests, I have tried to (statistically and psychometrically) ‘fish’ for any indication of bias possible with the data provided. Therefore, only if none of the analyses shows any possible effect would it be sufficiently plausible that the pass rates difference is due to true score differences.

Part 5 Results

In part 4 I listed the analyses in the order in which they were provided in the brief. For the sake of logic however, I will report them in a slightly different order addressing the three possible explanations (true difference, bias or error/chance finding) for the discrepancy in pass rates.

The first concern to address is whether the differences in pass fail rates between CC and NCC candidates is most likely due to general error or a chance finding.

5.1 Reliability analysis of both cohorts and calculation of standard errors of measurement (SEM) and 95% confidence intervals (95% CI).

Table 1: Descriptive statistics and reliability results

	mean	Standard deviation	Cronbach’s alpha	SEM	95%CI
Cohort 1	61.74	9.28	.837	3.75	7.34
Cohort 2	63.89	8.77	.788	4.04	7.91

The difference between the mean scores in both cohorts is small (roughly half an SEM), so it is less likely that one of the tests would have been biased and would have accounted for the difference in pass rates.

The reliabilities of the examinations of both cohorts is good enough for high-stakes testing according to the rules of thumb in the international literature (which use .80 as a minimum threshold). However, reliability in itself is not the most informative measure. It is an estimate of which part of the variance or standard deviation can be attributed to the variance due to differences in ability of candidates – so-called true score variance – and which part is measurement error. From the reliability and the standard deviation the standard error of measurement can be calculated which is a more concrete indication of the measurement error and can be used to determine the 95% CI around each candidate’s score or around the cut-off score. The SEMs and 95%CIs of the 2016.2 OSCE are similar to those found in many other OSCEs both in the undergraduate and post-graduate context. So in themselves these results do not support the assumption that the difference in pass rates would be attributable to measurement error.

5.2 Reliabilities of the NCC and CC group separately for both cohort and the calculation of all SEMs and 95%Cis for both cohorts and both groups of candidates. Plus item analysis and comparison between NCC and CC candidates to determine whether there are specific stations with high differences.

Table 2: breakdown of descriptive statistics and bias psychometrics per cohort and by background of candidates (CC – Caucasian; NCC – non-Caucasian)

		mean	Standard deviation	Cronbach's alpha	SEM	95%CI
Cohort 1	CC	66.63	7.73	.765	3.75	7.34
	NCC	55.37	7.03	.721	3.71	7.28
Cohort 2	CC	66.22	7.72	.737	3.96	7.76
	NCC	58.32	8.73	.768	4.20	8.24

Ideally reliabilities, SEM and 95% CIs are calculated at the level of interest. In this case – given the questions in the brief (bias, error or true score differences) a breakdown at the level of candidate group and cohort was needed. So, in order to examine the reliabilities more closely, I have analysed them for the performances of the CC group and the NCC group separately. It is clear that the 95% CIs are in a similar range for these subgroups as they were for the total groups in analysis 5.1 The difference between CC and NCC candidates in cohort 1 is 11.26 and the combined 95%CIs = 14.62, so the 95% confidence intervals overlap. In cohort number 2 the difference between both candidate groups is 7.9 and the combined 95% CIs = 16. This approach actually treats both means as individual data points in one distribution to determine whether the difference is large enough not to be caused by measurement error. Again the assumption of error being the cause of the pass rates difference is not supported by the findings: although there is a small overlap in 95% CI in cohort 1 and a slightly larger in cohort 2 this is not enough to explain the difference in pass rates.

5.3 Chi square test of candidate groups (NCC or CC) against passing or failing.

Another way of looking at the concern is to estimate the likelihood that the difference in pass rates has occurred by chance. This would be one of the alternative explanations for the findings (the other two explanation are: bias or real difference in ability).

From the data I was sent I have calculated the total scores by adding up the station scores (with double the value for stations 13, 14, and 15 in cohort 1, and 16, 17 and 18 in cohort 2) and divided them by 18. Using the cut-off score of 63% for cohort 1 and 64% for cohort 2. This led to the following table of pass and fail rates.

Table 3: Pass and fail rates breakdown by cohort and background of candidates.

		Cohort1	Cohort2	Total both cohorts	%-age
CC	Fail	16	22	38	(38/129)*100 =29.46%
	Pass	44	47	91	(91/129)*100 =70.54%
NCC	Fail	41	24	65	(65/75)*100 =86.7%
	pass	5	5	10	(10/75)*100 = 13.50%

Although there is still a marked difference in pass rates between the CC and NCC candidates 70.54% versus 13.50%) it is not as high as those rumoured (88% versus 6.8%). Using the results of my own calculations and compared with the calculations done by the College the following 2 x 2 table was constructed. (The numbers in brackets are the expected values for each cell).

Table 4: Contingency table of the pass and fail rates against background of candidates (pooled cohort 1 and cohort 2)

	pass	**fail**	***Marginal Row Totals***
CC	91 (63.87)	38 (65.13)	129
NCC	10 (37.13)	65 (37.87)	75
Marginal Column Totals	100	104	204

The results of the chi square analysis is: $X^2 = 62.0949$ which leads to a $p < .0001$.

This indicates that the likelihood that the found association between passing and failing and background of candidates is less than 0.01% (actually the likelihood would even be much lower as the critical value for X^2 with 2 df for a p of 0.0001 is 13.816, so 65.0418 is considerably higher. So this finding does not support the assumption that the difference is due to a chance occurance.

5.4 T-test between the two candidate groups (NCC and CC) for descriptive purpose only and compare with overlap of 95% confidence intervals.

Another way of looking at it is to see whether the difference in mean scores is likely to be coincidental or not. Given the numbers of candidates and given that the assumption of a normal distribution of total scores is plausible. I have used parametric statistics, in this case T-tests.

T-test NCC/CC cohort 1: T= 7.724, df = 104, $p<.0001$

T-test NCC/CC cohort 2: T = 4.449, df = 96, $p<.0001$

In both cohorts the difference between the mean scores of the CC group and the NCC group are significant (CC group scoring higher than the NCC group in both cohorts) which can be interpreted as (a proxy for) the likelihood of the differences being purely by chance is less than .01%. This finding also does not support the assumption that the pass rate difference is due to chance.

5.5 Determination of numbers of possible 'false-positive' and 'false-negative' results (score within a 95% CI around the cut-off score) and 'true-positive' and 'true-negative' results (scores outside the 95%CI around the cut-off score). Comparison between NCC and CC in both cohorts.

The cut-off score for cohort 1 was set to 63% and for cohort 2 set to 64% and from this I have used the SEM to construct a 95% CI around the cut-off score.

Table 5: 95% Confidence Intervals around the cut-off scores.

		Cut off score	SEM	95%CI	Lower bound	Upper bound
Cohort 1	CC	63%	3.75	7.34	55.66	70.34
	NCC	63%	3.71	7.28	55.72	70.28
Cohort 2	CC	64%	3.96	7.76	56.24	71.76
	NCC	64%	4.20	8.24	55.76	72.24

Using the 95% CI around the cut-off scores I have determined the proportions of 'true' negative results (those candidates whose score was lower than the cut-off score minus the lower 95% CI) and 'false' negative results (those with a score below the cut-off score but within the 95% CI) and the same for the true positives (above the upper 95%CI) and false positives (within the 95% CI).

Table 6: determination of the 'true' and 'false' passes and fails using the 95% CI.

		True negatives	False negative	False positives	True positives
Cohort 1	CC	5	11	22	22
	NCC	21	21	3	1
Cohort 2	CC	10	12	31	16
	NCC	14	10	3	2

If we were to look at only the true positives and true negatives in both cohorts the percentages passing and failing would be

Cohort 1

CC: 81.5% pass and 18.5% fail; **NCC:** 4.5% pass and 95.5% fail

Cohort 2:

CC: 61.5% pass and 38.5% fail; **NCC**: 12.5% pass and 87.5% fail

Or in total:

CC: 71.75% pass and 28.3% fail; **NCC:** 7.9% pass and 92.1% fail

So in conclusion, even if we only look at the true positive and true negative results there is a considerable disparity between the numbers of passing and failing candidates between the CC and NCC groups in both cohorts.

The results of analyses 5.1 – 5.5 make it unlikely that the difference in mean scores and the subsequent pass rates between CC and NCC candidates is due to general measurement error or a chance occurrence. Although in analysis 5.2 some overlap in 95% CIs was found this in itself is not enough to conclude that the difference is due to error. Therefore, the next steps are focussed on finding indications of more or less specific bias.

5.6 Calculation of difference between the mean scores on individual stations

The logical first step in evaluating the likelihood of this assumption is a more detailed analysis of the p-values or the mean scores per station. I have separated these out per group of candidates (CC versus NCC) and per cohort. With this I aim to examine whether there are specific stations that can be identified to be contributing to the difference in pass rates in an extreme fashion or whether it is a broader phenomenon. To explore this, I have simply subtracted the p-values of the NCC candidates from those of the CC candidates per station and for each cohort.

Table 7: *P-values of the 15 stations per candidate group for cohort 1*

Station number	p-value CC	p-value NCC	difference
1	69.33	57.14	12.19
2	69.29	55.15	14.13
3	65.22	58.88	6.33 (7)
4	61.90	52.92	8.98 (14)
5	61.52	53.63	7.89
6	62.45	49.13	13.32
7	65.71	52.30	13.42
8	65.45	53.98	11.48 (12)
9	63.90	57.05	6.85 (17)
10	70.43	59.25	11.17
11	62.31	54.41	7.90 (19)
12	61.00	47.67	13.33
13	68.69	53.82	14.87 (6)
14	74.40	64.69	9.71
15	67.29	54.10	13.19

The maximum difference is 14.87% and the minimum is 6.33%. In nine stations the difference is more than 10% (I have chosen 10% as an arbitrary cut-off) and in six it is less than ten percent. So, although there is variation in the extent to which stations contribute to the difference in pass rates, it is also important to notice all stations differences are in the same direction.

The differences in the shaded rows are the stations in which NCC examiners were involved, the numbers in brackets indicate the number of candidates for which this was the case. As is clear from the table, there is no clear tendency for these stations to have lower differences in p-values than the other stations, so there is not noticeable – measurable – influence of the origin of the examiner on the scores.

Table 7: *P-values of the 15 stations per candidate group for cohort 2*

Station number	p-value CC	p-value NCC	difference
16	71.06	55.62	15.44
17	62.89	56.65	6.24 (4)
18	72.71	65.22	7.49
19	73.50	66.60	6.90
20	65.38	58.72	6.66
21	69.71	64.43	5.28
22	58.80	47.78	11.02 (5)
23	65.07	53.25	11.82 (8)
24	62.73	54.68	8.05
25	65.63	52.66	12.97 (13)
26	63.73	57.00	6.73 (10)
27	72.63	70.64	1.99 (3)
28	62.27	61.08	1.18
29	61.94	58.96	2.98
30	57.39	49.01	8.38

In cohort 2, the maximum difference is 15.44% and the minimum is 1.18%, and so there is more variation in the differences. In this cohort, only 4 stations have a difference than 10% (again, chosen as an arbitrary cut-off) and eleven stations show a difference of less than ten percent. Similarly to cohort 1 there is variation in cohort 2 in the extent to which stations contribute to the difference in pass rates. In this cohort 2 all station differences are in the same direction as well and the stations with NCC examiners do NOT have lower differences than those without NCC examiners.

In both cohorts, the CC candidates outperform the NCC candidates on ***all*** stations regardless of whether there were exclusively CC examiners or a mix between CC and NCC examiners. So, I was unable to find any indication for a station specific bias or an examiner-background specific bias. This does not mean it could not have occurred, but as explained in the opening parts of this report, it is highly unlikely that any of the findings in analysis 5.6 would be sufficient to account for a discrepancy in pass rates of the magnitude found in the 2016.2 OSCE.

5.7 Calculation of the curriculum domain scores for both cohorts and comparison between the NCC and CC groups.

Another explanation could be that the NCC candidates have been disadvantaged by the inclusion of certain domains in the examination. For this, I have calculated the scores per domain for both cohorts and have compared these between the CC and NCC candidate groups. I have used T-tests as an indicator for the meaningfulness of the difference (or, if you will, a proxy for the likelihood that the difference is due to chance). I repeat that this is not performed with the intent to make any inferences as to whether ACEM examination in general would be biased or not against any group of candidates but merely as an indicator of the likelihood of the discrepancy between the pass rates of CC and NCC candidates being a chance occurrence.

Cohort 1:

In cohort 1 the domains ‘Medical expertise’, ‘Communication’, ‘Scholarship and teaching’, ‘Prioritisation and decision making’ and ‘Health advocate’ were examined in more than one station (“Leadership and management’, ‘Professionalism’, ‘Teamwork and collaboration’ only in one). Therefore I have compared the results of the CC candidates with those of the NCC candidates on those 5 domains only. To gauge the magnitude of the difference I have used T-tests. The results are presented in table 8.

Table 8: Difference between the p-values of the major curriculum domains in cohort 1

domain	**T**	**Degrees of freedom**	**P**
Medical expertise (k=21)	6.940	104	<.0001
Communication (k=8)	7.754	104	<.0001
Scholarship and Teaching (k=5)	6.223	103.966	<.0001
Prioritisation and decision making (k=4)	6.034	104	<.0001
Health advocate (k=2)	4.372	104	<.0001

Cohort 2:

In cohort 2 the domains ‘Medical expertise’, ‘Communication’ , ‘Scholarship and teaching’, ‘Prioritisation and decision making’ and ‘Health advocate’ were examined in more than one station (“Leadership and management’, ‘Professionalism’, ‘Teamwork and collaboration’ only in one). Therefore I have compared the results of the CC candidates with those of the NCC candidates on those 5 domains only. To gauge the magnitude of the difference I have used T-tests. The results are presented in table 9.

Table 9: Difference between the p-values of the major curriculum domains in cohort 2

domain	T	Degrees of freedom	P
Medical expertise (k=22)	3.972	96	<.0001
Communication (k=5)	4.105	96	<.0001
Scholarship and Teaching (k=6)	2.883	96	<.0001
Prioritisation and decision making (k=5)	4.857	96	<.0001
Health advocate (k=2)	1.152	96	.252

In both cohorts all domains but one, differences in mean scores were found (with the CC candidates scoring higher than the NCC candidates on all occasions). The likelihood for each of these to be the results of chance is low (<.01% probability); with the exception of the domain ‘Health advocate’ in cohort 2. This domain was only examined in two stations in cohort 2. Therefore it is extremely unlikely that it could account for the difference in pass rates of the total group could and that it would explain sufficiently the discrepancy in pass rates between CC and NCC candidates.

6 Conclusions

Determining whether an examination is biased purely from psychometric analysis is not easy. As explained in the opening parts of this report it is basically one equation with three unknowns:

1) the difference may be due to a real difference in ability between the candidate groups

2) the difference is due to a form of bias against one of the candidate groups

3) the difference is a one-off random finding and is due to error in the measurement

In summary, I was unable to conclude that explanation #3 is the most likely explantion. The reliabilities overall were good and so were those in the breakdown by candidate background and cohort. When I made a distinction in true and false positives and true and false negative (defined as either outside or within a 95% CI) and compared the ratios of only the true positives and negative for both groups, the pass-fail ratios were not dissimilar to those reported over the whole group. Finally the chi square and t-test made it likely that the difference is not a chance occurrence. In all, it is I therefore safe to conclude that explanation #3 (chance or error) can be ruled out.

The next explanation - a specific bias against NCC candidates- was explored by looking for markedly different performance of either certain stations, certain examiner groups or certain curriculum domains. In all cases, stations , examiner background and domains, the differences in performance were across the whole range and no specific stations, examiner background groups and/or domains could be identified that would sufficiently and plausibly account for the difference in pass rates. The content of the stations seemed to me – although I have limited expertise in emergency medicine – to be reasonable and not particularly Caucasian orientated. But I also assume that during the station construction process these issues have been addressed as well and the stations have been scrutinised for any such possible bias. Explanation #2 can therefore also be sufficiently be ruled out ***as the cause for the discrepancy in pass rates***. I have no way of determining whether any form of bias would have occurred in an individual situations – ***either against NCC or against CC candidates*** – but there is no indication that any form of bias big enough to account for the pass rate difference was present.

This leaves me with explanation #1, namely that the difference in performance between the CC and NCC group represents a difference in the ability the OSCE purported to measure. As I explained at the start of this document, there is no way to rule in or out any of the explanations with certainty; only their likelihood can be discussed. Therefore, given the combination of all analyses, I must conclude that explanation #1 is the most likely one for this examination.

7 General discussion and advice

There are many different ways in which high-stakes OSCEs are being administered around the world and it is fair to say that there will be numerous examples in which OSCEs like the ACEM's OSCE are used. However, I don't think that this is the most important question to address. Many assessment practices are based on beliefs and tradition. Following those particular examples and claiming that it is good practice because others are doing it may not be an optimal underpinning of quality. There is, however, a vast literature on OSCEs and there are some valuable lessons to be drawn from it. Without turning this report into a scientific paper I will highlight what I think to be the most pertinent findings in the literature with respect to the ACEM OSCE process.

- *Detailed checklists are not better than more global rating scales*

In its original form the OSCE relied on detailed checklists and short (5 minute) stations. The reason behind this was the belief that inter-rater reliability was the main cause for the unreliability in skills assessment. Very soon afterwards, however, it was found that *inter-case* reliability was the most important factor (labelled domain or content specificity) for unreliability and not the *inter-rater* reliability. The advice from these studies is to 'nest 'examiners within stations (as is usual practice with OSCEs) and not to use double marking. When more examiners are available, increasing the number of stations with one examiner each is more effective than having fewer stations with two examiners per station. The reason for the ACEM to have 2 examiners on certain stations may certainly add to the credibility of the process to its stakeholders, and show due diligence. This in itself can be a defensible reason for it, but psychometrically it is not necessary.

- *The most important aspect of validity of the OSCE is not the rubric but the examiner*

In any type of assessment there is subjectivity. Every type of assessment requires an evaluation of the performance/competence of candidates and therefore human judgement always plays a role. In multiple choice and other written types of examination the collection of the performance (candidate responses) is disconnected from the judgement processes (blueprinting, item selection, determination of pass fail scores, specific wording of the items, determination of answer keys, etc.), and the response collection and calculation of scores can even be done by computers. In any type of observation-based assessment (of which the OSCE is one) the collection of performance information and the judgement will have to go hand in hand. The examiner observes and interprets the performance at the same time. Such processes need expertise of the examiner. S/he does not only need to have sufficient expertise about the content of the station but also assessment expertise (what to look for, how to judge, how to score, what is acceptable performance, what is reasonable to expect of candidates, etc.). Research shows that this type of expertise develops much like diagnostic expertise develops (through the formation of scripts and automation or development tacit knowledge), which is logical because both diagnosis disease and diagnosing' dyscompetence' are both so-called diagnostic classification or categorisation tasks.
For the ACEM OSCE this implies that changes to the rubric should not be the first priority in development but a clear focus on examiner training to ensure that all examiners are sufficiently assessment literate for the OSCE. The rubrics as they are currently being used in the ACEM OSCE are of a type that do require sufficient assessment literacy or expertise. The literature suggest that more detailed rubrics support examiners with less experts/experience better. However, I would suggest to prioritise ensuring optimal examiner training (which I think is already part of the process) rather than any change to the rubrics.

- *Licensing examinations have to be such that they convince stakeholders*

Apart from its measurement characteristics, the OSCE examination is also important in reassuring stakeholders that those candidates who pass are most likely to be safe and independent practitioners. This often creates a discrepancy between what would be best-evidence based practice and what is acceptable practice. There is increasing pressure in the literature to not apply standard psychometric processes to OSCEs. The most obvious is the lack of plausibility of within-examination compensation (e.g., poor performance on an abdominal examination station can be compensated for with good performance on knee examination) which does not make sense in real practice. Alternative models are currently being studied (network psychometrics, IRT and even Bayesian probabilistic model) but they are not mainstream enough to be used in the extreme high-stakes and possibly litigious context as the ACEM's fellowship examination. So, there might be evidence in the literature that would suggest that current OSCE practice is not the best for the purpose but the alternatives would still be too new and therefore not suitable to convince stakeholders.

- *Standard setting is arbitrary but cannot be frivolous*

Standard setting is a process that seeks to dichotomise a continuous variable. The total percentage score can be any given value (between 0 and 100% usually) but has to be grossly subdivided into two categories ('pass' or 'fail'). There is no 'true' pass fail score, it is always the result of assumptions and decisions. In the borderline regressions method (BRM) the following assumptions need to be met:

- There volume of data is high enough to allow for a regression analysis
- The pass-fail score can be set at the level of the total examination allowing for inter-station compensation
- The scores on each station are numerical
- The global judgement about each candidate's performance is made by somebody with sufficient expertise
- The judgement is based on direct observation

In the case of the ACEM all assumptions are met; the date volume is sufficient, inter-station compensation is possible, the scores are numerical, the examiners have expertise and the judgements are based on direct observation. As such the BRM is defensible for the ACEM's OSCE and it can be explained well to all stakeholders.

In summary:

- The most likely explanation for the difference in pass rates between CC (70.5%) and NCC (13.5%) is a true difference in performance, and the other two assumptions (chance/error or systematic bias) are much less likely to the extent that no effects were found big enough to explain the difference.
- The ACEM's OSCE practice is similar to many other organisations' practice but this does not mean that it is therefore best evidence based. However, given the delicate context in which the ACEM has to operate, dramatic conceptual changes to the process would require careful planning and stakeholder engagement right from the start. This would be an entirely separate project.
- As for the details of the current practice with respect to the forms used, the examiners training and the use of the borderline regression method for pass-fail decisions these are all acceptable practice within the boundaries of the College's context.

-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-

Appendix 1: data conversion tables.

Cohort 1	Station scores file		
EXCEL COLUMN	Var number	Description in Excel file	Variable name in SPSS
A	1	Candidate identifier	
B	2	NCC or Caucasian graduate candidate	
C	3	Station 1 Examiner 1 identifier	EX1_ID_STAT1
D	4	Station 1 NCC or Caucasian Examiner 1	ORIG_EX1_STAT1
E	5	Station 1 Total station score (%)	SCORE_STAT1
F	6	Station 2 Examiner 1 identifier	EX1_ID_STAT2
G	7	Station 2 NCC or Caucasian Examiner 1	ORIG_EX1_STAT2
H	8	Station 2 Total station score (%)	SCORE_STAT2
I	9	Station 3 Examiner 1 identifier	EX1_ID_STAT3
J	10	Station 3 NCC or Caucasian Examiner 1	ORIG_EX1_STAT3
K	11	Station 3 Total station score (%)	SCORE_STAT3
L	12	Station 4 Examiner 1 identifier	EX1_ID_STAT4
M	13	Station 4 NCC or Caucasian Examiner 1	ORIG_EX1_STAT4
N	14	Station 4 Examiner 2 identifier	EX2_ID_STAT4
O	15	Station 4 NCC or Caucasian Examiner 2	ORIG_EX2_STAT4
P	16	Station 4 Total station score (%)	SCORE_STAT4
Q	17	Station 5 Examiner 1 identifier	EX1_ID_STAT5
R	18	Station 5 NCC or Caucasian Examiner 1	ORIG_EX1_STAT5
S	19	Station 5 Examiner 2 identifier	EX2_ID_STAT5
T	20	Station 5 NCC or Caucasian Examiner 2	ORIG_EX2_STAT5
U	21	Station 5 Total station score (%)	SCORE_STAT5
V	22	Station 6 Examiner 1 identifier	EX1_ID_STAT6

W	23	Station 6 NCC or Caucasian Examiner 1	ORIG_EX1_STAT6
X	24	Station 6 Total station score (%)	SCORE_STAT6
Y	25	Station 7 Examiner 1 identifier	EX1_ID_STAT7
Z	26	Station 7 NCC or Caucasian Examiner 1	ORIG_EX1_STAT7
AA	27	Station 7 Examiner 2 identifier	EX2_ID_STAT7
AB	28	Station 7 NCC or Caucasian Examiner 2	ORIG_EX2_STAT7
AC	29	Station 7 Total station score (%)	SCORE_STAT7
AD	30	Station 8 Examiner 1 identifier	EX1_ID_STAT8
AE	31	Station 8 NCC or Caucasian Examiner 1	ORIG_EX1_STAT8
AF	32	Station 8 Examiner 2 identifier	EX2_ID_STAT8
AG	33	Station 8 NCC or Caucasian Examiner 2	ORIG_EX2_STAT8
AH	34	Station 8 Total station score (%)	SCORE_STAT8
AI	35	Station 9 Examiner 1 identifier	EX1_ID_STAT9
AJ	36	Station 9 NCC or Caucasian Examiner 1	ORIG_EX1_STAT9
AK	37	Station 9 Examiner 2 identifier	EX2_ID_STAT9
AL	38	Station 9 NCC or Caucasian Examiner 2	ORIG_EX2_STAT9
AM	39	Station 9 Total station score (%)	SCORE_STAT9
AN	40	Station 10 Examiner 1 identifier	EX1_ID_STAT10
AO	41	Station 10 NCC or Caucasian Examiner 1	ORIG_EX1_STAT10
AP	42	Station 10 Examiner 2 identifier	EX2_ID_STAT10
AQ	43	Station 10 NCC or Caucasian Examiner 2	ORIG_EX2_STAT10
AR	44	Station 10 Total station score (%)	SCORE_STAT10
AS	45	Station 11 Examiner 1 identifier	EX1_ID_STAT11
AT	46	Station 11 NCC or Caucasian Examiner 1	ORIG_EX1_STAT11
AU	47	Station 11 Examiner 2 identifier	EX2_ID_STAT11

AV	48	Station 11 NCC or Caucasian Examiner 2	ORIG_EX2_STAT11
AW	49	Station 11 Total station score (%)	SCORE_STAT11
AX	50	Station 12 Examiner 1 identifier	EX1_ID_STAT12
AY	51	Station 12 NCC or Caucasian Examiner 1	ORIG_EX1_STAT12
AZ	52	Station 12 Examiner 2 identifier	EX2_ID_STAT12
BA	53	Station 12 NCC or Caucasian Examiner 2	ORIG_EX2_STAT12
BB	54	Station 12 Total station score (%)	SCORE_STAT12
BC	55	Station 13 Examiner 1 identifier	EX1_ID_STAT13
BD	56	Station 13 NCC or Caucasian Examiner 1	ORIG_EX1_STAT13
BE	57	Station 13 Examiner 2 identifier	EX2_ID_STAT13
BF	58	Station 13 NCC or Caucasian Examiner 2	ORIG_EX2_STAT13
BG	59	Station 13 Total station score (%)	SCORE_STAT13
BH	60	Station 14 Examiner 1 identifier	EX1_ID_STAT14
BI	61	Station 14 NCC or Caucasian Examiner 1	ORIG_EX1_STAT14
BJ	62	Station 14 Examiner 2 identifier	EX2_ID_STAT14
BK	63	Station 14 NCC or Caucasian Examiner 2	ORIG_EX2_STAT14
BL	64	Station 14 Total station score (%)	SCORE_STAT14
BM	65	Station 15 Examiner 1 identifier	EX1_ID_STAT15
BN	66	Station 15 NCC or Caucasian Examiner 1	ORIG_EX1_STAT15
BO	67	Station 15 Examiner 2 identifier	EX2_ID_STAT15
BP	68	Station 15 NCC or Caucasian Examiner 2	ORIG_EX2_STAT15
BQ	69	Station 15 Total station score (%)	SCORE_STAT15

Cohort 2	**Station scores file**		
EXCEL COLUMN	Var number	Description in Excel file	Variable name in SPSS
A	1	Candidate identifier	CAN_ID

B	2	NCC or Caucasian graduate candidate	ORIG_CAN
C	3	Station 16 Examiner 1 identifier	EX1_ID_STAT16
D	4	Station 16 NCC or Caucasian Examiner 1	ORIG_EX1_STAT16
E	5	Station 16 Examiner 2 identifier	EX2_ID_STAT16
F	6	Station 16 NCC or Caucasian Examiner 2	ORIG_EX2_STAT16
G	7	Station 16 Total station score (%)	SCORE_STAT16
H	8	Station 17 Examiner 1 identifier	EX1_ID_STAT17
I	9	Station 17 NCC or Caucasian Examiner 1	ORIG_EX1_STAT17
J	10	Station 17 Examiner 2 identifier	EX2_ID_STAT17
K	11	Station 17 NCC or Caucasian Examiner 2	ORIG_EX2_STAT17
L	12	Station 17 Total station score (%)	SCORE_STAT17
M	13	Station 18 Examiner 1 identifier	EX1_ID_STAT18
N	14	Station 18 NCC or Caucasian Examiner 1	ORIG_EX1_STAT18
O	15	Station 18 Examiner 2 identifier	EX2_ID_STAT18
P	16	Station 18 NCC or Caucasian Examiner 2	ORIG_EX2_STAT18
Q	17	Station 18 Total station score (%)	SCORE_STAT18
R	18	Station 19 Examiner 1 identifier	EX1_ID_STAT19
S	19	Station 19 NCC or Caucasian Examiner 1	ORIG_EX1_STAT19
T	20	Station 19 Total station score (%)	SCORE_STAT19
U	21	Station 20 Examiner 1 identifier	EX1_ID_STAT20
V	22	Station 20 NCC or Caucasian Examiner 1	ORIG_EX1_STAT20
W	23	Station 20 Total station score (%)	SCORE_STAT20
X	24	Station 21 Examiner 1 identifier	EX1_ID_STAT21
Y	25	Station 21 NCC or Caucasian Examiner 1	ORIG_EX1_STAT21
Z	26	Station 21 Total station score (%)	SCORE_STAT21

AA	27	Station 22 Examiner 1 identifier	EX1_ID_STAT22
AB	28	Station 22 NCC or Caucasian Examiner 1	ORIG_EX1_STAT22
AC	29	Station 22 Examiner 2 identifier	EX2_ID_STAT22
AD	30	Station 22 NCC or Caucasian Examiner 2	ORIG_EX2_STAT22
AE	31	Station 22 Total station score (%)	SCORE_STAT22
AF	32	Station 23 Examiner 1 identifier	EX1_ID_STAT23
AG	33	Station 23 NCC or Caucasian Examiner 1	ORIG_EX1_STAT23
AH	34	Station 23 Examiner 2 identifier	EX2_ID_STAT23
AI	35	Station 23 NCC or Caucasian Examiner 2	ORIG_EX2_STAT23
AJ	36	Station 23 Total station score (%)	SCORE_STAT23
AK	37	Station 24 Examiner 1 identifier	EX1_ID_STAT24
AL	38	Station 24 NCC or Caucasian Examiner 1	ORIG_EX1_STAT24
AM	39	Station 24 Total station score (%)	SCORE_STAT24
AN	40	Station 25 Examiner 1 identifier	EX1_ID_STAT25
AO	41	Station 25 NCC or Caucasian Examiner 1	ORIG_EX1_STAT25
AP	42	Station 25 Examiner 2 identifier	EX2_ID_STAT25
AQ	43	Station 25 NCC or Caucasian Examiner 2	ORIG_EX2_STAT25
AR	44	Station 25 Total station score (%)	SCORE_STAT25
AS	45	Station 26 Examiner 1 identifier	EX1_ID_STAT26
AT	46	Station 26 NCC or Caucasian Examiner 1	ORIG_EX1_STAT26
AU	47	Station 26 Examiner 2 identifier	EX2_ID_STAT26
AV	48	Station 26 NCC or Caucasian Examiner 2	ORIG_EX2_STAT26
AW	49	Station 26 Total station score (%)	SCORE_STAT26
AX	50	Station 27 Examiner 1 identifier	EX1_ID_STAT27
AY	51	Station 27 NCC or Caucasian Examiner 1	ORIG_EX1_STAT27

AZ	52	Station 27 Total station score (%)	SCORE_STAT27
BA	53	Station 28 Examiner 1 identifier	EX1_ID_STAT28
BB	54	Station 28 NCC or Caucasian Examiner 1	ORIG_EX1_STAT28
BC	55	Station 28 Examiner 2 identifier	EX2_ID_STAT28
BD	56	Station 28 NCC or Caucasian Examiner 2	ORIG_EX2_STAT28
BE	57	Station 28 Total station score (%)	SCORE_STAT28
BF	58	Station 29 Examiner 1 identifier	EX1_ID_STAT29
BG	59	Station 29 NCC or Caucasian Examiner 1	ORIG_EX1_STAT29
BH	60	Station 29 Total station score (%)	SCORE_STAT29
BI	61	Station 30 Examiner 1 identifier	EX1_ID_STAT30
BJ	62	Station 30 NCC or Caucasian Examiner 1	ORIG_EX1_STAT30
BK	63	Station 30 Examiner 2 identifier	EX2_ID_STAT30
BL	64	Station 30 NCC or Caucasian Examiner 2	ORIG_EX2_STAT30
BM	65	Station 30 Total station score (%)	SCORE_STAT30

Cohort 1	**Curriculum domain scores**		
Id in Excel file	Var num in SPSS	description in Excel file	Variable name in SPSS
A	1	Candidate identifier	CAN_ID
B	2	NCC or Caucasian graduate candidate	CAN_ORIG
C	3	Station 1 Medical expertise (1)	MED_EX_1
D	4	Station 1 Medical expertise (2)	MED_EX_2
E	5	Station 1 Scholarship & teaching	SC_TEACH_1
F	6	Station 2 Medical expertise (1)	MED_EX_3
G	7	Station 2 Medical expertise (2)	MED_EX_4
H	8	Station 2 Scholarship & teaching	SC_TEACH_2
I	9	Station 3 Medical expertise (1)	MED_EX_5
J	10	Station 3 Medical expertise (2)	MED_EX_6
K	11	Station 3 Communication	COMM_1
L	12	Station 4 Medical expertise (1)	MED_EX_7
M	13	Station 4 Medical expertise (2)	MED_EX_8
N	14	Station 4 Communication	COMM_2
O	15	Station 5 Medical expertise (1)	MED_EX_9
P	16	Station 5 Medical expertise (2)	MED_EX_10
Q	17	Station 5 Prioritisation & decision making	PRI_DEC_1
R	18	Station 6 Prioritisation & decision making	PRI_DEC_2
S	19	Station 6 Medical expertise	MED_EX_11
T	20	Station 6 Leadership & management	LEAD_MAN_1
U	21	Station 7	MED_EX_12

		Medical expertise	
V	22	Station 7 Communication	COMM_3
W	23	Station 7 Communication	COMM_4
X	24	Station 8 Communication	COMM_5
Y	25	Station 8 Health advocacy	HEALTH_AD_1
Z	26	Station 9 Medical expertise (1)	MED_EX_13
AA	27	Station 9 Medical expertise (2)	MED_EX_14
AB	28	Station 9 Scholarship & teaching	SC_TEACH_3
AC	29	Station 10 Communication	COMM_6
AD	30	Station 10 Medical expertise	MED_EX_15
AE	31	Station 10 Professionalism	PROF_1
AF	32	Station 11 Medical expertise	MED_EX_16
AG	33	Station 11 Scholarship & teaching	SC_TEACH_4
AH	34	Station 12 Medical expertise	MED_EX_17
AI	35	Station 12 Communication	COMM_7
AJ	36	Station 12 Health advocacy	HEALTH_AD_2
AK	37	Station 13 Medical expertise	MED_EX_18
AL	38	Station 13 Prioritisation & decision making	PRI_DEC_3
AM	39	Station 13 Communication	COMM_8
AN	40	Station 14 Medical expertise	MED_EX_19
AO	41	Station 14 Prioritisation & decision making	PRI_DEC_4
AP	42	Station 14 Teamwork & collaboration	TEAM_COLL
AQ	43	Station 15 Medical expertise (1)	MED_EX_20
AR	44	Station 15 Medical expertise (2)	MED_EX_21
AS	45	Station 15 Scholarship & teaching	SC_TEACH_5

Cohort 2	**Curriculum domain scores**		
Id in Excel file	Var num in SPSS	description in Excel file	Variable name in SPSS
A	1	Candidate identifier	CAN_ID
B	2	NCC or Caucasian graduate candidate	CAN_ORIG
C	3	Station 16 Medical expertise	MED_EX_1
D	4	Station 16 Prioritisation & decision making	PRI_DEC_1
E	5	Station 16 Teamwork & collaboration	TEAM_COLL_1
F	6	Station 17 Medical expertise	MED_EX_2
G	7	Station 17 Prioritisation & decision making	PRI_DEC_2
H	8	Station 17 Communication	COMM_1
I	9	Station 18 Medical expertise	MED_EX_3
J	10	Station 18 Scholarship & teaching	SCHOL_TEA_1
K	11	Station 19 Medical expertise (1)	MED_EX_4
L	12	Station 19 Medical expertise (2)	MED_EX_5
M	13	Station 19 Scholarship & teaching	SCHOL_TEA_2
N	14	Station 20 Medical expertise	MED_EX_6
O	15	Station 20 Scholarship & teaching	SCHOL_TEA_3
P	16	Station 21 Medical expertise (1)	MED_EX_7
Q	17	Station 21 Medical expertise (2)	MED_EX_8
R	18	Station 21 Communication	COMM_2
S	19	Station 22 Medical expertise	MED_EX_9
T	20	Station 22 Prioritisation & decision making	PRI_DEC_3

U	21	Station 22 Communication	COMM_3
V	22	Station 23 Medical expertise (1)	MED_EX_10
W	23	Station 23 Medical expertise (2)	MED_EX_11
X	24	Station 23 Prioritisation & decision making	PRI_DEC_4
Y	25	Station 24 Medical expertise	MED_EX_12
Z	26	Station 24 Prioritisation & decision making	PRI_DEC_5
AA	27	Station 24 Leadership & management	LEAD_MAN_1
AB	28	Station 25 Medical expertise (1)	MED_EX_13
AC	29	Station 25 Medical expertise (2)	MED_EX_14
AD	30	Station 25 Medical expertise	MED_EX_15
AE	31	Station 26 Medical expertise (1)	MED_EX_16
AF	32	Station 26 Medical expertise (2)	MED_EX_17
AG	33	Station 26 Scholarship & teaching	SCHOL_TEA_4
AH	34	Station 27 Medical expertise	MED_EX_18
AI	35	Station 27 Communication	COMM_4
AJ	36	Station 27 Health advocacy	HEA_ADV_1
AK	37	Station 28 Medical expertise	MED_EX_19
AL	38	Station 28 Scholarship & teaching	SCHOL_TEA_5
AM	39	Station 29 Medical expertise (1)	MED_EX_20
AN	40	Station 29 Medical expertise (2)	MED_EX_21
AO	41	Station 29 Scholarship & teaching	SCHOL_TEA_6
AP	42	Station 30 Medical expertise	MED_EX_22
AQ	43	Station 30 Communication	COMM_5
AR	44	Station 30 Health advocacy	HEA_ADV_2