INDEPENDENT REVIEW OF THE ASSESSMENT PROCESSES OF THE ROYAL COLLEGE OF ANAESTHETISTS

Professor John C. McLachlan

Contents

Executive summary

This Review was commissioned as a consequence of various errors in the delivery of the Royal College of Anaesthetists (RCoA) examination processes in 2021 and the resulting wish of the RCoA to have a fully independent review of assessment within the College. It was carried out by means of interviews of key RCoA and Faculty members, to which past candidates and doctors in training (both anaesthetists and intensivists) were also able to contribute. Surveys were also circulated to relevant stakeholders. Extensive documentation was made available for review, along with previous surveys of examiners. A rapid review of relevant literature was also conducted.

An additional strand was added to the Review to explore the culture and behaviour of examiners. This was largely carried out by interviews and study of current and previous surveys of examiners.

The overall impression gained was that there have been serious divisions between the College and the candidates. Some candidates were plainly resentful and suspicious of College actions and described these in negative terms. RCoA communication with the candidates had generally been viewed as poor.

However, while understanding how such strong feelings could arise in the course of a high stakes and very challenging assessment process, I found no evidence that the RCoA was, in fact, uncaring or unconcerned about candidates, although at times the exams assessment processes had been severely challenged by the requirement to respond to the pandemic. College staff and officers also at times felt unfairly blamed by some of the criticism levelled at them. Exams are also internally reviewed every five years, and changes and recommendations are made in the reviews.

I hope that this independent Review will address some of these mutual misunderstandings. I will recommend closer integration of candidates and doctors in training in College processes and would urge those individuals to engage as fully as possible in making their views and concerns known within the College. I would also encourage College staff and officers to make every effort to share their perceptions and challenges with actual and prospective candidates.

Equally, I was struck by accounts of the stress that the current examinations system places on candidates. There were heartfelt accounts of major and long-lasting disruption to personal and family life, of stress related to study and the assessments, especially failures, of trainees being deterred from pursuing careers in anaesthesia, and even, on rare occasions, of thoughts of self-harm. Some of my recommendations are directed at clarifying the purposes of the exams, and reconsideration of pass rates and standards, thereby reducing the stresses imposed on candidates.

I also noted a number of vulnerabilities and inconsistencies in the processes across the four major sets of assessments and have made recommendations directed at these.

Two major errors occurred in 2021, one relating to the character limit in place for Constructed Response Questions, and one relating to the inappropriate release of some personal data to candidates. These derived in large part from the stresses of converting assessments to online versions due to the pandemic and are unlikely to occur again in these forms. In general, the actions taken by the RCoA to mitigate the effects were as reasonable as was possible under the circumstances. I also considered the very low pass rate observed in the FFICM OSCE, which, while not an error *per se*, caused considerable concerns to candidates. I note the data available were difficult to access in the available timeframe. While the decision to allow the results to stand was not an unreasonable one, I have made recommendations on how to address a comparable issue arising again.

I recommend full involvement of the Exam Team (the staff employed by the College under the leadership of the Head of Examinations) at all stages of the exams, so that their expertise informs all stages of the assessment process. While the four sets of assessments properly have different purposes and constituencies, I recommend that they should have consistency of best practice in methodology (including standard setting and adjustments). This would streamline examiner training and ensure consistency of standards across all the assessments. I recommend that clear lines of governance and accountability should exist across the College and Faculty assessment processes.

I devote some consideration to the purposes and standards of the assessments. 'Knowledge' appeared across all components of the assessments, despite some of the assessment formats being inappropriate for reliable assessment of this domain. 'Skills' had some representation, although the OSCEs (optimal for skills assessment) also contained inappropriate knowledge elements. `Capability in the workplace' and `patient safety' were represented implicitly rather than explicitly.

Clarifying the purposes of the assessments should also clarify the appropriate standards. It is nowhere suggested that a large proportion of the current training workforce in anaesthetics and intensive care or pain medicine is incompetent. Fail rates in RCoA exams should be commensurate with capability in the workplace. I make recommendations on standard setting and research which should help address the issue of appropriate fail rates.

Concerning assessment formats, for the written examinations, I recommend moving to Single Best Answer MCQs as rapidly as possible. The OSCE currently contains inappropriate pure 'knowledge' stations which should be removed, and current practice is not sufficiently flexible in the interpretation of candidate responses. The Structured Oral Examinations (SOEs) are plainly valued by some candidates and examiners, and I recommend moving the invaluable skill of 'clinical reasoning', seen in the SOE at its best, into the OSCE format. This would reduce total testing time and ease standard setting. A significant saving in testing time and cost would thereby result, benefitting both candidates and the operation of the assessment processes.

With regard to assessment delivery platforms, I recommend a continuous review of the software market. Current manual data handling approaches in the RCoA are vulnerable to error.

In relation to standard setting, I recommend consistency of practice across all four sets of assessments. While the format of the exams is changing, I recommend use of the Hofstee compromise method to ensure that there are no rapid and inappropriate resulting swings in fail proportions. I recommend the engagement of psychometric expertise to aid College deliberations and propose several research projects which would materially empower appropriate standard setting and assessment design.

I recommend that the College has an identified Equality, Diversity and Inclusion Lead and that Equality Impact Assessments are conducted with regard to significant changes in assessment.

Finally, concerning examiner culture, I recommend that formal assessment forums such as WhatsApp and Slack, and settings such as 'Call Over' discussions, should be for professional exams related matters, avoiding irrelevant material, and that diversity and inclusion remain key aspirations for examiner recruitment.

The number of recommendations I make does not suggest that the RCoA is unique in having challenges relating to the complex process of assessment. Other medical professional organisations who invite independent external scrutiny would also be likely to receive recommendations for change.

1. Introduction

1.1 Background to the Review

The Royal College of Anaesthetists (RCoA) examinations are part of the postgraduate medical training programme in the UK and, for this purpose, are approved by the GMC. They comprise four sets of examinations: the Fellowships of the Royal College (FRCA) Primary and Final examinations, the Fellow of the Faculty of Intensive Care Medicine (FFICM) examinations and the Faculty of Pain Medicine (FPM) examinations. All four exams are delivered by the processes, systems and staff of the RCoA. An internal review of the FRCA examinations process is currently underway, exploring, *inter alia*, the timing, test design and standard setting methods. Written assessments employ Multiple True False (MTF), Single Best Answer (SBA) and Constructed Response (CRQ) items. Practical and Oral assessments include Objective Structured Clinical Examination (OSCE) and Structured Oral Examination (SOE) methods.

During the time of the pandemic, the RCoA switched its assessments to online versions to avoid face-to-face contact, in line with Government guidance. This was the case with almost all professional assessments during this time. Much of this process was successful, and all the assessments were delivered in one form or another. However, two particular challenges arose in 2021. In one case, each successful candidate in the FRCA Final received details of all the successful candidates via e-mail. In another, the Constructed Response (CRQ) section of the FRCA Final only permitted entry of a very limited number of characters, which did not allow candidates to answer the questions adequately.

The RCoA issued video apologies for these issues, and stakeholder meetings were organised to receive commentary on these events. From these, it emerged that there were serious concerns on the part of candidates, in particular, relating not only to the particular errors, but also to RCoA communications in general, and the administration and structure of the examinations as a whole.

As a consequence, it was decided to initiate a completely independent Review of the RCoA assessment processes. While this included consideration of the particular errors that had occurred, it was to be much wider in scope, covering the IT systems and infrastructure, the resources available, including staffing, the roles played by staff and examiners, the processes and delivery methods of the assessments, item banking, standard setting, and communication and contingency planning. The remit covered the FFICM, FPM and FRCA examinations. The author was commissioned to carry out this Review, and the agreed methodology is summarised in the next section.

An additional strand was added to the Review to explore the culture and behaviour of examiners.

On entering into the review process, I indicated that I would be cognizant of the issue of change management in response to recommendations I might make. Conscious that the College contains the Faculties of Intensive Care Medicine and Pain Medicine, I therefore resolved to make generic recommendations applying to all aspects of the examination programmes. If an assessment element within the suite of exams already accords with the recommendation, well and good: if it does not, then the recommendation can be read as applying to that element.

1.2 Methodology: how the Review was carried out

There were four main approaches to conducting the Review.

1.2.1 Interviews

Virtual interviews via Teams or Zoom were carried out with key individuals involved with the RCoA and the assessment processes. These included Chairs and Deputy Chairs of the relevant

organisations, Exam Team staff, examiners, candidates and representatives of doctors in training. An open invitation was extended to stakeholders to contact me directly, and a number of individuals made use of this opportunity. All interviews were conducted with guarantees of confidentiality. Interviews were semi-structured and with regard to the main Review, occupied one hour. Hand-written notes were taken during the meeting, which I expanded via speech-to-text transcription immediately after each meeting, and finally summarised in a constructed version in Microsoft Word. Analysis was by a modified grounded theory approach. Data saturation was reached, at which point no new themes were emerging, towards the end of the review process. Direct quotes, either from interviews or from surveys or other documentation, are placed in italics and quotation marks.

Sixty direct invitations were issued to individuals and organisations to contribute, and in the end, over 70 individuals were spoken with, including some representing larger groups of individuals.

1.2.2 Document review

A considerable variety of documents including meeting minutes, previous reviews, correspondence, surveys and spreadsheets were made available to me. These were reviewed and key points were fed back into the interviews and the written summaries.

1.2.3 Surveys

A customised survey was sent to examiners, Anaesthetists in Training, College Tutors, TPDs and Regional Advisers, and to the FICM. The survey was open for a three-week period. Over 150 responses were received, divided into several categories. These free text comments were reviewed thematically and informed both the subsequent interviews and this Review.

1.2.4 Rapid literature review

A rapid literature review was conducted around various themes, including equality and diversity, and technical psychometric issues. Relevant references are provided as Endnotes in this document, and the full literature review grid is available on application to the author.

1.2.5 Examiner culture

A review of examiner culture was conducted via Zoom or Teams interviews, from a representative list provided by the College. Each interview was scheduled over 30 minutes. Questions about examiner culture were also added to the structured interviews to obtain a much wider pool of respondents. I also had access to several surveys relating to examiner culture carried out in recent years.

2. General impressions

During this Review, it became clear that there was a significant divide between candidates and the RCoA, most frequently expressed as mistrust and suspicion on the part of candidates towards the RCoA, but also on occasion, on the part of RCoA representatives towards groups of candidates. I believe that exploring this divide frankly through the medium of this Review, uncomfortable though it may sometimes be, is essential to restoring trust on both sides, to the long-term benefit of candidates, the College, and ultimately patients. I also believe that with a better understanding on both sides, an appropriate professional relationship can indeed be developed between candidates and the RCoA.

It is essential to appreciate just how important the exam processes are for candidates. During the Review, I heard of candidates who felt that their lives were put on hold during the time of preparing for the exam, with friendships and family regularly set aside and important events such as marriages

and holidays delayed. Frequently, candidates were in the period of the life cycle where families are young, and career and life decisions are being made and implemented. Despite these important factors, the exams still came to dominate their lives. The burden of the exams was, of course, added to that of working in an environment which is highly demanding at the best of times, and which was exacerbated by the special demands of the pandemic.

During the interviews, a number of interviewees became emotional while recounting their feelings and experiences. For candidates who had failed an assessment, issues of lowered self-esteem were reported regularly. They reported that exam outcomes were also used as a status marker by others in the workplace, not just in terms of formal career progression but also as informal estimation of their personal value. These feelings were also present in the surveys, and on a number of occasions, mental health concerns were expressed.

These impacts were not confined only to the time of exams or to unsuccessful candidates. One candidate who had been successful at the first attempt in all the exams felt that s/he was unable to restore damaged relationships which had suffered during the time of the examination.

The assessment process is also demanding in financial terms, both for the cost of the exams themselves and also for the associated travel and accommodation issues. The limited number of diets per year exacerbates the timing problems for candidates, automatically building in delays. In the event of failure, or another problem arising, there may be an extended wait before a new attempt is possible.

The limited number of diets also makes the exams impossible to sit discreetly: the requirement for study leave means that it is well known who is sitting the assessments at each opportunity, and the outcome, whether pass or fail, rapidly becomes clear to others. Estimates of self-worth seemed particularly challenging when candidates had received positive workplace-based assessments that turned out to be at odds with outcomes of the assessment process. Further reference will be made to the issue of concurrent validity later in this Review.

Concerns were also expressed by some individuals about the availability of materials to support preparation for the assessment, and about the feedback offered to candidates who had failed an assessment.

Communication from the RCoA and Faculties, particularly FICM, was generally perceived as poor. Candidates reported that they found out late about important events, and that information was delayed and unclear. Individual and social media communications were valued more than generic feedback such as via a website.

High levels of mistrust and suspicion, either of the College as a whole or a faculty, were voiced by several past candidates. Quotes along the lines of *"I don't trust or value them anymore"* were received from several past and present candidates.

Conversely, some comments from College examiners and committee members represented dismay at the attitudes of candidates, particularly as expressed through some of their representative groupings. RCoA staff and officers sometimes felt they were unfairly maligned and accused of lack of care for candidates or incompetence and idleness, when, in fact, a great deal of work and care had been invested in introducing the pandemic assessment processes.

However, having spoken very frankly with a number of College officers and staff, I feel it would be fair to say that accusations that College and Faculty officers and staff were uncaring, incompetent, or malignly motivated were not a true representation of the situation. 'The College' is not some

monolithic entity. It is made up of a relatively small number of men and women who over the time of the pandemic were wrestling with a variety of novel problems and unexpected crises, against the backdrop of a pandemic which negatively affected their own working relationships and practices, along with, as with everyone else, their lives and those of their loved ones. College staff and officers, too, frequently expressed real emotional upset, sometimes in response to being blamed for events which were outside their control. Yet in general, the errors and problems that arose should not detract from the fact that the RCoA was able to respond to the pandemic by introducing widespread responses and changes which in many circumstances worked well and would be appropriate grounds for congratulation under normal circumstances.

Mistakes were certainly made in the stress of circumstances. Some of these resulted from work pressures, especially under the novel circumstances of the pandemic, and the resulting emotional stresses on those implementing the examination processes. These can be addressed by reflective analyses of particular events. However, others were the visible manifestation of certain underlying structural problems which were likely to result in errors of one kind or another, even if the exact form was not foreseeable. Moreover, some of these problems remain and await chance events to bring about more problems in the future. Structural changes will be recommended that should reduce the risk of these events taking place.

The best possible outcome of this Review would be to bring the College and the candidates back together in a mutually beneficial and co-operative way. To this end, changes will be recommended to the administrative structures and to the assessment processes which, I believe, will make the outcomes more proportionate to the trust that society, through the NHS, has placed in all of the candidates. If there is a mismatch between the evaluation of their ability and the outcomes of the assessment, then such a discrepancy must be explored to identify how it has come about as a matter of urgency.

3. Problems that arose in 2021

Two significant assessment errors arose in the course of the 2021 assessment period. I was asked to look at each of those, and to this, I have added the issue of the pass rate of the October 2021 FFICM OSCE, which was plainly of concern to many respondents.

3.1 An inappropriate character limit was imposed in the RCoA Final CRQ

3.1.1 What happened and what was done?

During the Final CRQ examination on 14th September 2021, candidates sitting the Final FRCA CRQ examination found that the number of characters that could be entered in the free text boxes was limited (and inadequate to the required task). The platform provider TestReach paused the exam to resolve the issue and resumed the exam when the restriction was lifted. This meant that there were very significant variations in candidates' experience and their ability to answer the questions.

The agreed resolution was that the September 2021 Final FRCA Written examination result would be based solely on the MCQ component. Candidates who did not pass the September 2021 Final FRCA Written examination would not have that sitting counted toward their total number of attempts and would be eligible for a free resit in the March 2022 examination, with the validity period being extended if required. An apology was issued by the President.

3.1.2 Why did it happen?

At this point, the assessment was delivered by TestReach using the 'Quick Start' process, under the stress of making the assessments online in a short period of time. In this, the RCoA provided the

paper, and TestReach built it. This did not allow the Exam Team access to the backend of the programme, so they were unable to see the settings in place. Initially, the CRQ items were built using an essay format, which allowed unlimited characters, but subsequently, TestReach changed this to a character-limited format, without the Exam Team being made aware of this. There had also been an understandable 'examiner' desire to limit the amount candidates wrote, but along the communication channels, there seems to have been confusion over 'word limit' and 'character limit'.

There were, therefore, multiple and confusing lines of both responsibility and communication within the College, and to and from TestReach, leading to a misunderstanding about the length, and even the existence, of a character limit.

3.1.3 How appropriate were the actions taken?

Giving a public account of the nature of the error and apologising were appropriate initial steps. Under the circumstances, basing the outcomes on the MCQ component was the correct decision. The CRQ and MCQ both attempt to measure the same construct – knowledge. The MCQ avoids the issue of assessor variability, and the MCQ results are, if anything, more reliable than the CRQ results. The actions that were taken seem to have been the best available in the circumstances. Offering a free resit and not counting the attempt were not strictly necessary in psychometric terms but represented a recognition of the challenge to candidates that had arisen during the exam and were also appropriate.

3.1.4 How can such an error be avoided in the future?

Plainly, some of the challenges arose from the stress of the pandemic, with the engagement with TestReach, in particular, being made under the challenges of setting up an online system from scratch. Moving away from the 'Quick Start' aspect of the platform will help. But a more important consideration is that decisions on issues such as word or character limits were made without the Head of Examinations being aware of them or being in direct communication with the platform provider.

I recommend in Section 4 that the Exam Team, and particularly the Head of Examinations, must be involved in a central role in all the assessment processes, rather than the Exam Team acting as an 'auxiliary' or 'service' department, only carrying out decisions made by others. Not only is this good practice in itself, but it also reduces markedly the possibility of errors such as this re-occurring. I understand discussions have already taken place with TestReach, to whom some share of responsibility for communication errors seems to accrue, to clarify lines of communication. I make a recommendation in Section 4 about clear lines of accountability which is relevant to this issue.

However, there is an additional issue about employing several different item formats in the same exam. This reduces the length of each subtest, posing a challenge to reliability, but also complicates the whole process, increasing the risk of errors. I have recommended in Section 7.1.5 that the written format should rely entirely on single best answer MCQs for the foreseeable future.

Particularly if my Recommendations are followed, I do not expect that this error will arise again. There is an issue, however, about the quality and volume of the data available at the time decisions were being made. It would have been helpful if, for instance, the correlation of previous candidates' MCQ and CRQ scores had been available, to confirm the best course of action available. I address this issue in Section 8.1.

3.2 The inappropriate release of candidates' results

3.2.1 What happened and what was done?

This arose from a 'mailmerge' error, in which 237 candidates who had passed were each also informed of the name, honorific and College Reference Number of other candidates who had been successful. No resolution of this issue was practicable.

3.2.2 Why did it happen?

Due to the stress of operating in the pandemic (and also in dealing with the consequences of the character limit error described above), an individual omitted to close the 'Edit' field in the 'Mail Merge' programme before the results were sent out. This resulted in the details of all successful candidates being sent out to each recipient.

3.2.3 How appropriate were the actions taken?

This was properly reported to the Information Commissioner's Office. A video apology was made, again appropriately, and a meeting was organised at which stakeholders could provide their input and express their concerns. The error did not lend itself to remediation, in that once the candidate information had been sent, it could not be 'unsent'.

3.2.4 How can such an error be avoided in the future?

I would describe this as a stress-related error, rather than a process error, and as long as staff are not placed under such extreme stress again, I would class it as unlikely to recur in this form. However, I also note that the Exam Team appear frequently to be placed under time stress, due to the exam timetable, and this should be taken into account in workforce and timetable planning.

3.3 The October FFICM OSCE pass rate.

3.3.1 What happened and what was done?

Although this was not an error in the same sense as the above, I was told that there was a particular area of concern relating to the discrepancy between the pass rate (28%) at the October 2021 FFICM OSCE and previous iterations of this exam (where it is normally around 67%). There was also a discrepancy with the corresponding SOE pass rate, which was about 59%; normally this records a pass rate of about 69%, very similar to the OSCE.

This is very surprising, even in the unusual circumstances of the pandemic. Normally candidates perform reasonably consistently across time, and even across formats: strong candidates tend to be strong even across several different assessment methodologies. Cohorts of candidates also tend to perform fairly predictably.

The occurrence of this discrepancy obviously caused extensive dismay among candidates, who raised concerns both formally and informally.

Good practice when such a discrepancy occurs obviously involves a check on the administration of the exam – were the correct results recorded, was the standard setting procedure correctly carried out, how were items selected, was there an influence of new examiners, including sampling videos of OSCE stations, and so on. I understand such a process took place and could not identify an evident error in the exam administrations. Most items had been used previously without difficulties arising.

Focus then turned to the candidate pool – were candidates more likely to be resitting the assessment in this administration of the exam than in other administrations? However, the data available were difficult to access in the available time frame.

As a result of these processes, no clear reason to vary the outcome was observed, and the results were, therefore, allowed to stand. A free resitting was offered to candidates who had failed.

3.3.2 Why did it happen?

As indicated above, no single clear cause was identified. A common comment from respondents was that training itself had been affected seriously by the pandemic since both trainers and trainees were working in very unusual environments, focused on a much more limited range of conditions than normal.

3.3.4 How appropriate were the actions taken?

It is likely that the decision taken to allow the OSCE results to stand was the best decision that could be justified at the time, in the light of the information and advice available.

Should similar circumstances arise again, it might be worth considering, in collaboration with the regulator, the possibility of test linkage, especially where there is overlap between the material covered in different elements of the testing process.

The point about the impact of the pandemic is a substantial one. All over the UK, training programmes and assessments were disrupted by the pandemic at every level. In most environments, steps were taken to mitigate the impact of these disruptions on candidates' progression, even accepting some loss of information about candidate standards.

It was clearly signalled to me by several candidates that failure in exams made them question their career choices, not just with regard to career progression, but sometimes also to the discipline itself, and even with regard to medicine, asking why they should put themselves through this again. People lost to the career, or false negatives in an exam, are societal risks to patient safety.

3.3.4 How can such a situation be avoided in the future?

In Section 10, I will recommend that during the period when the assessment structures of the RCoA are changing markedly, the Hofstee standard setting methodology should be implemented as a 'reality check' on the existing standard setting methods. This will be particularly desirable in the event of a marked discrepancy between new outcomes and those observed previously, in the absence of identifiable causative factors.

Recommendation 1: If a marked discrepancy from previous results is observed without an identifiable causative factor, the Hofstee compromise method be employed as the primary standard setting approach.

4. Governance and committee structures

A theme which emerged a number of times throughout the Review was that of the governance of the assessment process. The Exam Team were widely, indeed fulsomely, praised both in interviews and surveys. However, there was also a sense that their expertise was not used to its full potential. In particular, the idea that they were sometimes considered as a 'service department' emerged a number of times while it is clear that their expertise merited regarding them full and equal partners in the assessment process. Naturally, the examiners and the relevant committees across the RCoA have subject matter expertise. But subject matter expertise is not the same as assessment expertise, particularly with regard to: (a) technical issues relating to assessment such as standard setting; and (b) practical matters of delivery of assessment. I heard of occasions in which the Exam Team heard rather late in proceedings of proposed initiatives which may not have been practical in the time scale available. I believe that the Exam Team should form a more integral part of the assessment

process at all stages, and that, in particular, the Head of Examinations should be a full member of all committees involved in decision making about assessments.

Recommendation 2: The Head of Examinations or colleagues to which they delegate responsibility, be a full member of all relevant assessment committees, including Faculty bodies, in a decision-making role on a par with other committee members.

I am aware that it has been suggested that the terms of reference of the Primary Examinations Review Group (PERG) and the Final Examinations Review Group (FERG) of FRCA be amended to create two exam delivery groups reporting to an Exam-Strategy Group. This is not one of my recommendations, but I appreciate that it would ease the task of implementing common best practice across the various assessments, which is one of my recommendations (as described in Section 5).

The 'them versus us' perception by some candidates with regard to their relationship with the RCoA was of concern and is associated with issues relating to communication with candidates. There is already trainee representation on various bodies within the RCoA, but this representation did not seem to have resolved the challenges. Besides committee membership where exam content is discussed or candidate identity and outcomes are at issue, it would be helpful to have candidate and doctors in training representation at as many stages of the assessment process as possible, and for those trainees to have a role in proposing the means of communication to candidates.

Recommendation 3: If candidate and doctors-in-training representation in an advisory capacity on all RCoA committees engaged in assessment be assured, these representatives play a key role in supporting communication with candidates.

It was sometimes unclear to me how the governance structures operated across the various exams and faculties. The absence of clear governance is likely to have contributed to the error regarding the character limit described in Section 3.1.

Recommendation 4: Clear lines of governance and accountability should exist across all the College and Faculty assessment processes.

5. Consistency of best practice

A variety of differing practices have arisen across the Primary and Final assessments, and across the Faculties, for instance concerning standard setting. Evidently, there is different content and intentionality across these various assessments. But in my view, the variations in practice do not derive from these appropriate differences, and they create weaknesses and divergence from best practice. They may also inhibit transitions from one role to another, for instance, if examiners move from Primary to Final examinations. Common practice would bring about labour cost benefits and improve the defensibility of assessment processes. I will comment further on standard setting in Section 10 of this Review, but this would be perhaps the prime example of where best practice should be observed consistently across all exams run under the aegis of the RCoA.

Similarly, a unified approach to item writing, question banking and exam analysis would offer many advantages, including the possibility of sharing assessment materials and expertise where appropriate. I will comment in Section 8.1 on assessment platforms and software, but a shared approach would empower all of the contributors to the assessment programme.

Since the Faculties rightly value their independence, sharing common best practice should not be seen as a 'takeover' by the main RCoA. Differences in content and intention between the four sets of assessments will still exist. Rather, it should be seen as the sharing of best practice and the opportunity for economy of effort by reducing unnecessary differences which may have arisen historically.

Recommendation 5: Identification of best assessment practice on issues such as standard setting, item writing, and item banking be followed by the application of these practices consistently across all the assessments run under the aegis of the RCoA.

6. What are the assessments trying to measure?

I have placed this 'philosophical' question ahead of detailed considerations of operational details such as assessment formats, as it seems to be both essential and currently rather unclear to me within the RCoA at present. Is, for instance, the Primary viewed as an entry-level exam for Finals, or is it an independent entity in its own right? Are the exams meant to be career-determining progression steps or markers of more general excellence? How do they relate to the (currently largely formative) Work-Place Based Assessments (WPBAs) and Structured Learning Events taking place in the clinical environment?

And at an even more basic level, which domains are they attempting to assess? Knowledge is the only clear domain which was named to me at all stages of the assessments, although the methodologies used for assessing this were not always the most appropriate ones. Skills were much less frequently mentioned, although some of the assessment strategies were clearly designed to measure skills rather than knowledge. And 'patient safety' or 'capability in the workplace' were implicit rather than explicit in the detailed planning of assessments.

In addition to purpose, there also seemed some uncertainty as to the appropriate level of the assessments. Candidates, and indeed some examiners, viewed aspects of the exams as inappropriately detailed. Candidates also expressed concern about the levels of knowledge required and questioned whether examiners could demonstrate this knowledge without the foresight of seeing the questions in advance. The assessments were not viewed as matching career development with concerns raised that the standard of the exam was too high for the mid-term level of training as it was perceived to be set at post-CCT level. An examiner questioned whether it was necessary for candidates to know the molecular structure of a particular drug off the top of their head. And questions were raised about the relationship with career stage and the proportion of candidates working appropriately in clinical practice yet failing the exams. The high fail rate in the October FFICM OSCE described above is a case in point. Many of these doctors were already taking on responsibilities greater than those they would normally be expected to because of the pandemic.

Another challenge is the relationship between the task faced by candidates in retrieving information from memory at short notice, and that faced by examiners who have advance knowledge of the questions. Many plainly conscientious examiners described to me how they prepared before and during the exam period, appropriately reviewing the questions and possible answers in advance. But candidates then felt that they were asked questions without any opportunity for preparation. Candidates are faced with the task of retrieving information on unexpected topics from memory without much in the way of cueing[1], which is very much more difficult.

One line of enquiry that I adopted was related to the success rates of candidates in the assessments compared to their performance in the workplace, since most of the examiners were also tutors and trainers. The most common view, in terms of screening tests, was that 'false negatives' were more common than 'false positives': in other words, that candidates who were perceived as good in the clinical workplace nonetheless failed the exams, and that this was more common than candidates perceived as weak in practice nonetheless passing the exams.

Obviously, it is not my role to define the purpose of the assessments in a medical Royal College. But such definitions should be readily available, in close consultation with trainers and candidates at all

levels. Patient safety should, of course, feature prominently in such definitions. The relationship of the assessments to career progression should be carefully considered and clearly expressed.

There is also a clear perception from some that the current barrier is too high (see previously indicated concerns about the pass rate). The 'concurrent validity' research recommended in Section 11.3.1 of this Review should clarify the relationship between clinical performance and exam success. In general, the proportion of candidates who fail exam assessments should accord, in general terms, with the proportion who require further development in the workplace, and this does not currently seem to be the case. Some respondents mentioned consultants in the workplace, indicating that exam failure was not a good reflection on how good a trainee was in the clinical environment. This concurrent validity review should also help with addressing the stress experienced by candidates as they prepare for the assessments, and this, in turn, may ease the stress on the professional practice of anaesthesia. There were reports of anaesthetists dropping out of anaesthetic training and working at staff grades because of the exams being so stressful.

In Section 11, I recommend research studies to establish the concurrent and construct validities of the RCoA assessments. In Section 10.3, I also recommend the use of the Hofstee standard setting method, particularly while the format of items in the assessments is changing. Use of Hofstee has a benefit that it is unlikely to vary markedly from estimations of the general capabilities of candidates in practice.

Recommendation 6: A clear statement of the intended purpose of the assessment be drawn up for all the assessments, and that the results of benchmarking by doctors-in-practice at various levels be used to inform the standard setting procedures.

7. Assessment formats

7.1 Written assessments

7.1.1 Multiple True-False Items (MTFs)

MTFs have been part of the RCoA assessments for years. However, currently these are not the most favoured form of selected response item. In what is widely regarded as the definitive guide to writing selected-response items, Case and Swanson[2] indicate that "the various forms of true/false items are the worst". Previous studies[3] have also suggested that MTF items are less discriminating and less reliable. Since the practice of medicine is inherently ambiguous, MTF formats may promote an emphasis on 'true-but-trivial' information and may even promote the learning of misinformation[4]. However, the biggest challenge to the use of MTFs is probably that of standard-setting. One recent paper[5] lists 27 different methods of standard-setting MTF assessments, and a major problem is the requirement for correcting for guessing. While I have technical concerns about the corrections for guessing currently employed by the RCoA, these are outwith the scope of this Review, especially since TFs are being phased out and are the subject of separate technical discussions with the Exam Team. Here I will merely say that the phasing out of MTFs cannot take place soon enough, and if it is possible to accelerate the schedule for their removal, then this should be done.

7.1.2 Single Best Answer MCQs

The RCoA is gradually eliminating MTFs in favour of SBA MCQs with four alternatives (ie a key and three distractors). It should be noted that the occasional use of three or five alternatives can ease *item* writing and does not significantly harm the psychometric properties of the items. While

Extended Matching Items may perform even better than SBAs in some instances, they are more difficult to construct due to the difficulty of finding homogenous long-scale alternatives.

When using Angoff standard setting methods, SBAs do not require guessing correction as long as the Angoff values do not fall below 10/n, where *n* is the number of alternatives. Review of a recent RCoA Finals spreadsheet indicates that the Angoff values are all well above 10/n, so no further guessing correction will be required.

Good SBA MCQs can test all levels in the Knowledge Domain of Bloom's Revised Taxonomy[6], apart from Creativity. This, however, requires insightful development of the item stems to avoid simple tests of recall. In general, adding more clinically relevant information to the stem may mean it takes slightly longer to read, but this is then compensated by speeding up the process of arriving at the correct answer[7]. There is no reason to distinguish between MCQs on the basis of their length: all MCQs should have equal weighting when using Angoff standard setting approaches.

7.1.3 Constructed Response Items

CRQs are a part of the Final RCoA exams, which is also rather unusual these days in the assessment of knowledge in high stakes professional assessments. In general, the view is that high stakes assessments should emphasise Reliability, and the advantage of Selected Response items, such as MCQs, is that there is no component of assessor variance, and, therefore, the task of achieving and measuring reliability, if a good assessment platform is used, is simplified[8].

In my view, a further influential factor is that of cost. MCQs require advance time investment in writing and standard setting, but scoring and analysing them is then low cost, and the costs do not scale with candidate numbers. Once an MCQ has been written, it can be delivered to 50 candidates or 500 candidates at the same cost. But with CRQs, the cost 'scales up' directly when it comes to candidate numbers. Five hundred candidates cost ten times as much to mark as 50, with the added problems of assessor fatigue when large numbers are involved. Schuwirth and van der Vleuten[9] indicate that:

> Open-ended questions should be used solely to test aspects that cannot be tested with multiple-choice questions. In all other cases the loss of reliability and the higher resource-intensiveness represent a significant downside. In such cases, multiple-choice questions are not less valid than open-ended questions.

MCQs may also show lower 'differential attainment' problems than CRQs, especially for candidates for whom English is not the first language[10]. Moving to MCQs may, therefore, help address differential attainment gaps between candidates of different cultural backgrounds.

7.1.4 Very Short Answers (VSAs)

There is current interest in the use of Very Short Answer (VSA) items[11], where a clinical stem is answered as free text by the candidate, often in the form of one or two words. The advantage of these is that they can largely be computer-marked, with only unexpected answers being referred to an assessor. VSAs remove the effect of cueing (although real-world situations may indeed include multiple cues). However, while VSAs are of interest, they are more 'difficult' in that mean scores will be much lower than corresponding MCQs. Standard setting must be extensively re-thought under these circumstances. I, therefore, do not currently recommend their introduction to the RCoA assessments, as I am in favour of assessments utilising one format only (see Section 7.1.5). Perhaps a watching brief should be kept on how practice develops in this area.

7.1.5 Written assessments: conclusions and recommendations

The reliability of an assessment is generally in direct proportion to the number of items[12] (for instance, the Standard Error of Measurement decreases as the test gets longer). On this basis, the

longer the test the better. Moreover, using a single test format means that standard setting is simpler, and all analyses are more powerful. Item banking is also simpler, and item comparisons are empowered for issues such as test equating. On these considerations, I make the following recommendation.

Recommendation 7: The written components of all RCoA assessments be based on Single Best Answer MCQs, rather than Multiple True False or Constructed Response Questions, as soon as possible.

7.2 Practical and oral assessments

7.2.1 Objective Structured Clinical Examinations (OSCEs)

OSCEs are a well-established and widely used assessment tool in high stakes healthcare assessment, and as such, it is appropriate that they feature in the RCoA assessments. However, it was clear from both examiners and candidates that there were also significant reservations about how they were currently employed.

First, it is clear that there is an explicit intention to measure aspects of the Knowledge Domain within the RCoA OSCE format, and, indeed, there were previous examples of 'unmanned' stations where knowledge measurement is the sole purpose. But the recent Ottawa Consensus statement[13] is clear:

> OSCEs should be used to test clinical and communication skills. This was the original design intention … and still remains a key principle underlying the use of this assessment format. Globally, OSCEs have become the assessment method of choice for testing clinical and communication skills in an examination setting.

Of course, knowledge **could** be tested during an OSCE, but this is an expensive and inefficient way of measuring it. Moreover, if there are 'knowledge only' stations, then there are fewer 'skills' stations, and this means that the reliability of the important skills outcomes is reduced. Mixing two different test constructs means that standard setting also becomes much more complex.

There is also a challenge with 'knowledge' components in an OSCE, in that there is generally a correct answer, which can be shared between candidates if it is the same throughout the exam period. If the questions are varied between circuits and days to avoid this possibility, then this creates a problem of item variance, when candidates effectively sit different exams.

As the 'written' exams are transferred to online versions, the knowledge components can be tested through these means. Images, audio and video clips, and animations are all possible in the online version. OSCE stations should solely be used for the testing of clinical and communication skills.

A number of examiners and candidates also indicated undue requirement for candidates to employ the **exact** wording used in the examiner's guide in order to gain credit. This, allegedly, was to promote consistency and hence reliability. But the power of the OSCE in achieving reliability lies more in the range and number of stations and examiners, rather than in the reliability of any individual station. A revision of the examiner guides so that a degree of discretion is given to examiners to exercise judgement within the context of the marks scheme and checklist (as takes place in the SOE) will be recommended in Section 7.2.3. It may be valuable to extend the length of each station very slightly to accommodate further testing of the skill of clinical reasoning.

There are already some simulation stations, and these can readily be enhanced, and OSCEs focused on true skills, as is appropriate. However, removing the knowledge component to the written exams

does create some spare capacity with the current OSCE framework, and this will prove helpful when considering the outcomes for the SOE.

7.2.2 Structured Oral Examinations (SOE)

While reservations about the SOE, particularly in terms of examiner variability and the scoring process, were frequently expressed by respondents, there was plainly also a great deal of affection and positivity addressed to this comfort-zone assessment method. It was felt that the SOE could better show the difference between a borderline candidate and a weak one and made it possible to explore more thoroughly than an OSCE. It was also noted that the SOE was expensive to deliver.

However, there are reasons why this kind of structured oral exam is not widely used in high stakes healthcare assessments. Assessor variance is a major risk and is sometimes the biggest single cause of variance in outcomes for observed assessments[14]. This general principle was reinforced by descriptions of examiner style provided both by candidates and fellow examiners, with examiners described variously as more or less distant or friendly, or indeed helpful, with some candidates reporting being nudged towards the correct answer.

Another major risk in oral exams is item variance, where the use of different questions on different occasions creates different paths through the assessment for candidates. Again, this general principle was illustrated by specific candidate comments. One candidate felt they had had a more difficult exam than a colleague had had two days later, with both agreeing that one set of questions seemed easier.

The SOE is too short for assessor and item variance to be compensated by the number of 'stations'. It is unlikely that four or even six questions on knowledge or clinical reasoning will reach the necessary reliability level.

Concern was expressed about the use of a very short scoring scale (0, 1 or 2) which still managed to generate the maximum difference on occasion. Attempts are being made to extend the scoring scale, and/or use borderline regression, but trials have not so far produced clear benefits. This is possibly because borderline regression requires use of a check list score and a global judgement. At present, the scoring system in the SOE is effectively a global judgement, and adding a second global judgement thereby regressing the results is unlikely to be informative.

Because of the challenges of assessor and item variance, 'oral examinations' are currently not well accepted by external bodies such as the GMC. High stakes assessments should be *valid* and *reliable* within the technical meanings of these terms. The validity of the SOE is unexplored and the reliability will be low almost by design. Moreover, the reliability is currently not quantified by the RCoA in any consistent manner. Observation and feedback to examiners is not the same thing as data on examiner consistency, of kinds which could readily be quantified by use of a good software platform.

7.2.3 Practical and Oral Assessments: Conclusions and Recommendations

My conclusions spring from my view that it has not been clearly defined by the RCoA or the Faculties what the various assessments are attempting to measure. Knowledge testing is attempted in MTFs, SBAs, CRQs, OSCEs and SOEs. Skills testing is not clearly articulated as the most appropriate purpose of the OSCEs.

It is clear that there are attempts to test knowledge in the SOE, where it is just as expensive and unreliable to test as it is in the OSCE. But from the qualitative comments I received from both candidates and examiners, it seemed to me that what they were actually valuing most in the SOE

was the exploration of **clinical reasoning**. This is much more akin to a skill than knowledge, and as such lends itself well to precisely the kind of assessment found in OSCEs. This seems to me to offer both a solution to the challenges posed by the SOE, and an opportunity for resource savings, particularly in time costs for examiners, but also exam time for candidates.

I have noted above that removing knowledge content from the OSCE creates spare capacity in this framework. What I am proposing is that the best of both worlds can be obtained by incorporating the current SOE clinical reasoning content, minus 'knowledge', into the existing OSCE.

A clinical reasoning station of this kind can draw on existing material in the examining bank, and in the OSCE framework, would only require one examiner per station. It would be valuable to employ domain marking as part of this process, in conjunction with a global judgement score, which would empower good and consistent standard setting.

I believe that adoption of these strategies of removing knowledge testing from the OSCEs and the incorporation of the valuable clinical reasoning virtues of SOEs into a modified OSCE pattern can lead to considerable economies of scale for the assessment process as a whole. Each 'SOE' OSCE station can run effectively with one assessor, and it may be possible to almost half the time cost to candidates and assessors, compared to the current practice. The current OSCE assessment of knowledge is inappropriate and should be removed, but as this creates space in the OSCEs, then the virtues of the SOE can be folded into the OSCE structure. Reducing the overall time envelope of the assessment would also reduce the accommodation costs for examiners, which are currently considerable. Since I understand that the College does not profit from the assessments, cost savings can be passed on to candidates or invested in improving the quality of the exams, for instance by financing the essential research projects I have described in Section 11, or by investing in the Exam Team.

The benefits of reducing the length of the assessments are not merely financial, extensive though these might be. They would also include reducing examiner fatigue, currently described as significant by many respondents.

Recommendation 8: The 'knowledge' stations and components currently present in the OSCE be moved to the written papers.

Recommendation 9: The examiner guides for the OSCE encourage the exercise of judgement within the context of the marks scheme in interpreting candidate responses when they are clearly on the right lines.

Recommendation 10: (a) The 'knowledge' components of the SOE be moved to the written papers and; (b) the clinical reasoning skills components of the SOE be placed within the format of the OSCE.

8. Assessment delivery

8.1 The software platform and the need for better data

Currently, the RCoA employs two platforms, TestReach and Practique, for various parts of assessment delivery. A recent review has explored the use of other potential platforms without identifying a single optimal platform for the processes. I understand that, of all the offerings, TestReach is still currently viewed as being most capable of delivering the required online proctoring. This is a rapidly advancing area of technology, and while there may be no single ideal solution at the moment, it is quite possible that this will be available soon. I will recommend continued monitoring of the market. There are a variety of metrics and capabilities that would be

valuable. It cannot be emphasised enough that the current systems of recording and analysing assessment data are highly vulnerable.

Most platforms contain a question bank, in which items are stored along with their usage and performance (Facility), and metadata such as the topic, item developer, area of the curriculum, etc. This allows exams to be generated semi-automatically, and structured feedback to be generated for candidates. Use of a common platform across all assessments would allow Primary, Finals and the Faculties to share items and data where desired.

The need to record demographic data and match it to assessment outcomes is described in Section 11.2.

For each exam, Cronbach's α or KR20 and the mean discrimination index should be automatically generated, with means and standard deviations, along with data on comparative examiner performance where relevant. Outcome and pass lists should, of course, be generated automatically.

For each item, data such as the Facility, discrimination index or point bi-serial, Horst PKI and time required to answer can be generated. Items can be tracked over time, and any sign of item drift can be reviewed for evidence that the item has been compromised (particularly important when online delivery is being used). Internal 'collusion detection' capability such as the Harpp-Hogan Index can also be included.

I understand the RCoA currently operates an internal bespoke portfolio system. Such in-house systems tend to appeal to organisations when they are being developed, with the slightly mythic view that they 'will do exactly what we want'. The difficulty arises in maintaining and modifying such in-house systems. By contrast, commercial providers offer solutions with predictable fixed costs, and providers are often very helpful in adding specific capabilities to their platforms. I would suggest keeping an eye on this market also.

Recommendation 11: Commercial assessment platforms are kept under constant review, with a short- to medium-term view to implementing a single platform across all assessments delivered by the RCoA team. This will empower the collection and analysis of both biometric and psychometric data.

8.2 Online delivery of assessments

I understand that the decision to move to online at-home delivery of the written assessments, with candidates working from self-selected venues, has effectively already been made, and I accord with this decision, which is also generally supported by both interview and survey respondents. Candidates who had been based overseas during the exam period welcomed online assessments particularly. Online at-home delivery requires proctoring, and this, therefore, limits the choice of providers. However, as indicated elsewhere, this market is undergoing rapid change, and regular monitoring of available options is required.

For the current OSCE and SOE assessments, there were mixed views, but there appeared to be a majority preference, both on the part of examiners and candidates, for a return to face-to-face assessment. Comments were made about the importance of body language in interpreting complex situations, and some kinds of experience were viewed as not being readily replicable online.

8.3 Unfair means

With the use of online written exams, challenges will undoubtedly arise with regard to exam security (and, to a limited extent, these were already reflected in the surveys). It may be necessary to engage with appropriate software, as part of the final assessment platform, to measure item exposure and

the possibility of collusion. However, in my experience, 'cheating' is rarely a benefit to candidates since it imposes a cognitive load on candidates which is greater than the benefit. But concerns may arise among capable candidates, who tend to fear the possibility that others may benefit inappropriately from cheating, and these capable candidates may raise the issue more forcefully than might be expected. Just as medical practice encounters the 'worried well', so assessment may encounter what I might call the 'bothered brilliant': high performing candidates with anxiety about preserving and demonstrating their excellence. The use of proctoring and the need for observation during assessments may cause privacy concerns among candidates, and the Equality Impact Assessments recommended in Section 12 will need to address reasonable adjustments for candidates. In Section 8.1, I have mentioned the desirability of software that can detect collusion and item compromise.

8.4 OSCE sites

Occasionally in interviews and surveys, the requirement to attend London for live-presence exams was raised as expensive and problematic for candidates. I have recommended the combination of the OSCE and SOE experiences in Section 7.2.3, but there will still be face-to-face experiences, and I suggest (short of a formal recommendation) that an additional site for face-to-face assessments might be considered at a location which is relatively easy to reach, and at which OSCE-suitable venues might be available. Manchester might be a suitable location and is used by some Royal Colleges for this purpose.

9. Examiner selection, training, support and development

9.1 Style guide

A unified style guide should be adopted for all assessments. The NBME guide to MCQ writing[15] is widely used for this purpose. Similarly, the conversion of SOE items into clinical reasoning OSCE stations should be guided by a common format across all the assessments, so that items may be shared where necessary.

Recommendation 12: Common style guides for MCQs and OSCEs be introduced across all assessments delivered by the RCoA team.

9.2 Training

I have indicated in Section 4 that unified processes for all examiners should be instituted, so that common standards are observed, and examiners can move readily from one exam to another, confident that consistent, familiar and high standards are present in each. Aspects of the training can be recorded and made available online, again ensuring consistency from year to year and allowing face-to-face training to focus on high-yield activities (such as addressing the anxieties of new examiners). I have recommended in Section 13 on examiner culture that any inappropriate communications and comments made in face-to-face and online settings be eradicated by appropriate training and guidance.

Recommendation 13: Common high quality training materials and approaches be employed across all assessments delivered by the RCoA team, and a culture of continuous reflection and improvement be maintained.

9.3 Selection and recruitment

In interviewing examiners, I was struck by two things. The first was the high value all examiners had placed on educational issues throughout their careers, with examiners reporting long-standing interests in education and the second was how much they had valued the experience of being an examiner once they had joined. One examiner described as the best thing s/he had done, not just in their career but in their life. Plainly, in terms of forming support networks and developing as a

person and a clinician, being an examiner has been a very positive experience. Because of these positive features, I believe that being an examiner should be promoted as widely and inclusively as possible to those who might be interested and that the requirements for becoming an examiner should be re-considered to ensure that as wide a pool as possible meet the necessary conditions. Many of the issues with examiner culture described elsewhere in this Review would disappear if the pool of examiners were sufficiently diverse.

Recommendation 14: The inclusive recruitment of examiners should continue to be promoted by the College, as a personal, professional and societal benefit, and the requirements for becoming eligible to be an examiner are reviewed with a view to inclusion.

10. Standard setting

A number of comments and recommendations will be made with regard to standard setting.

From discussions with those involved in standard setting, there appeared to be a slightly exaggerated faith in standard setting *methodologies*, not entirely justified by evidence. For instance, in one document I reviewed, Borderline Regression was described as the 'gold standard'. It is certainly not that, and represents, rather, merely a favoured method of standard setting. The true value, the 'gold standard' of a standard setting method can only really be explored retrospectively, and to this end, I have recommended (in Section 11) various research projects I believe should be undertaken by the RCoA. The key question is "what are the assessments attempting to measure", and this is addressed in Section 6 of this Review.

It is not that the 'wrong' methodologies are used since they are ones in common usage in very many healthcare assessment settings, but that they are used (a) inconsistently across the different RCoA assessments, (b) unjustifiably with regard to guessing corrections and the subtraction of the standard error of the mean, and (c) with greater confidence in the outcomes than is justified.

10.1 Consistency

It would be a major advantage if common methods were used across all the RCoA assessments, so that experience can be shared, training rationalised and perfected, and software use simplified. Here again, use of an appropriate assessment platform would greatly ease calculations, limit costs to the RCoA, and reduce the risks of errors. Currently, standard setting takes place by use of idiosyncratic Excel spreadsheets, inherited from now retired colleagues, which exam personnel operate but do not, by their own account, fully understand. Indeed, it is a very difficult task to follow someone else's Excel spreadsheets, especially when these are not properly labelled or annotated, and it is something of a testament to the Exam Team that they have managed to employ them successfully so far, with no errors emerging. But this is a vulnerable (and time consuming) way to proceed. By contrast, a good assessment platform will have standard setting methods such as Angoff and Borderline Regression in-built, so that calculations and outcomes are available automatically and immediately on the conclusion of the assessment.

Consistency of methodology does not mean consistency of outcomes. Different assessments may still have different focuses and standards. But a common approach, with common software, will save money and time, and reduce risks.

Angoff and Borderline Regression remain the most popular methods of standard setting in high stakes healthcare assessments, and there is no reason to abandon their use in the RCoA. The changes I have recommended for the written and practical exams will indeed in time facilitate their more effective use. However, during changes to assessment practice, use of a compromise method can be invaluable, and I have recommended this in Section 10.3.

Recommendation 15: A unified approach to the use of, and training in, standard setting be employed across all the assessments delivered by the RCoA team.

10.2 Difficult to justify procedures

These are two-fold. First, subtraction of a standard error of the mean (SEM) from the cut score in the Final written exam is most unusual as a professional practice – in fact, as far as I can ascertain, it is unique in high stakes postgraduate medical assessments in the UK. It is more common to *add* an SEM to calculated cut-scores. Subtracting an SEM has the effect of making the exam 'easier' to pass, and the rationale I was offered was that without this, the fail rate might be 'too high'. This is interesting as it suggests that there is a *gestalt* perception of what the fail rate 'ought to be', and I will discuss this further below. However, if the current application of the Angoff method is giving an cut score which is somehow 'too high', this should be addressed directly, rather than by an additional and unusual procedure.

The second area of concern is the handling of guessing corrections. These are used inconsistently across the four sets of exams, and this is an example of an area where consistency of best practice (Recommendations 5 and 15) should be employed.

As described in Section 7.1.2, guessing corrections are not necessary for SBAs as long as the Angoff value is above 10/n, where n is the number of alternatives. Guessing correction is necessary for MTFs but should be used in a consistent manner. Phasing out MTFs should proceed as rapidly as possible, as indicated in Recommendation 7. Any perceived need for a 'guessing' correction will disappear when the written exams become all SBA in format, and the Angoff procedure is applied consistently.

Recommendation 16: The practice of subtracting a Standard Error of Measurement should cease, and guessing corrections be employed consistently across the RCoA exams, and should cease as soon as MTFs have been phased out in favour of SBAs.

Since cessation of the practices of subtracting an SEM and removing 'correction for guessing' is likely to have an impact on pass rates, I will recommend in Section 10.3 that the Hofstee compromise method be employed until the proposed changes to the assessment structures are well established. I do not recommend the practice of adding an SEM to the cut scores either. Once the changes to assessments have worked their way through, and standard setting methods are operating with success as compared to the results of the research recommended in Section 11, it is not necessary to make the assessments more difficult. I would reiterate that patient safety can be harmed by 'false negatives' as well as 'false positives'.

10.3 Use of compromise methods

Whenever major changes in either assessment methodology or assessment principles are undertaken, there is a risk that application of the same standard setting methods as used previously will be inappropriate. It is, therefore, an established practice to have a back-up or safety net process in place to respond to marked swings in outcomes. For instance, the GMC-ordained national Medical Licencing Assessment will employ Angoff standard setting methodology and the Hofstee compromise method[16] as a backup if the first results are too discordant from previous expectations. The Hofstee Method is a hybrid of prospective methods based on test items and retrospective methods based on test takers. It is relatively straightforward and cost-effective to implement. I will discuss the use of Hofstee in the context of the RCoA separately with the Exam Team, but a good exam platform will be able to implement Hofstee as part of its inbuilt capabilities.

Recommendation 17: The Hofstee Compromise Method be used to standard set assessments while the recommended changes to the RCoA assessment structures are taking place in order to ensure there are no inappropriate swings in pass/fail rates resulting from these changes.

11. Psychometric advice and the research agenda

11.1 Psychometric advice

Currently, there is mathematical expertise present in the Exam Team, and, to a limited extent, among examiners who have undertaken psychometric training programmes. However, I do not believe that this reaches the level of sophistication which is necessary to inform the RCoA deliberations. For instance, Generalisability Theory [17] (G Theory) analyses would inform decisions about how many OSCE stations are required to reach good yet economical levels of reliability and enable quantification of assessor and item variance. Similarly, mention was made in the current internal review of assessment of the possibility of using Item Response Theory[18] (IRT) as part of the analytic armoury of the RCoA. But neither G Theory or IRT are straightforward to implement, and my analysis is that the current Exam Team are too pressed for time to be able to add them to their repertoire. Equally, there are several sophisticated modelling techniques which could add real value to RCoA analyses.

It would, therefore, be invaluable to engage with professional psychometric assistance in some way. This need not be done by the addition of a full-time member of staff with the requisite expertise to the College staff roster. Psychometric expertise could be 'bought in' for specific purposes on regular occasions throughout the year. Perhaps, in line with one suggestion I have seen in minutes of an assessment review, expertise could be shared with another Royal College. Nonetheless, I am firmly of the view that further help is not merely desirable, but in the medium term, essential.

Recommendation 18: A professional psychometric capability be added to the College to support the Exam Team and exams' committees.

In addition, there are a number of possible research projects relating to (a) differential attainment and (b) validity, which would empower the design and analysis of RCoA exams. In particular, studies of validity would feed back in an invaluable way to the process of standard setting. I will explore these aspects in the following sections.

I was informed that RCoA has a research arm, which undertakes health research. There are many similarities between health research and educational research. Assessment is indeed in many ways a 'screening test' just like those used in clinical practice, where the 'gold standard' is patient outcomes. There may be possibilities to use research expertise within the RCoA to explore some of these issues. Alternatively, the RCoA might fund its own research, or seek external research funding (perhaps through NIHR) to pursue these issues. Recommendations for specific research projects are made in the following Sections, and the results should contribute to a culture of continuous reflection and improvement across all the College assessments.

11.2 Differential attainment

In higher levels of medical training, there remain significant issues of differential exam and career success by certain demographics, particularly with regard to gender, ethnicity and educational background, and internationally, this is true of anaesthesiology as well [19].

11.2.1 Gender

With regard to gender, the international literature recognises differences in outcomes between male and female candidates[20], and indeed this issue was the subject of a recent issue of the British Journal of Anaesthesia[21]. In a UK study in 2009, female candidates also seem to have performed less well than males in the RCoA assessment processes[22], and any potential differential performance of males and females in the RCoA exams would be of some concern. In general, such differential performance may be due to a genuine difference in the underlying property being measured or to an artefact of the assessment methodology. In comparable high stakes medical assessments, it is not uncommon

for women to perform better than men in professional assessments[23] and, for instance, for female surgeons to have better outcomes than male surgeons, particularly for female patients [24,25,26]. The proportion of female OR team members may also be significant for patient outcomes, with cooperation and safe practice being better when the male-female ratio is 50% or less [27]. These considerations would suggest that there is not an underlying construct difference in male and female performance in operating room settings. Indeed, if anything they suggest that increasing the proportion of female anaesthetists could bring patient benefits. Therefore, differential attainment of female candidates in the RCoA assessment processes should be quantified and addressed as a matter of importance. This can be explored through analysis of Differential Item Function[28], which, in turn, requires data on gender and ethnicity to be made available, perhaps through the Trainee Database. But collecting current and ongoing data on the performance of candidates by gender is essential, and is part of Recommendation 19 below.

11.2.2 Cultural background

Less clear-cut recent evidence is available on the performance of other demographic factors such as ethnicity, English as a first or second language, place of education and training, and other potential factors. However, older evidence relating to anaesthesia training[29] suggests there may be differential attainment in evidence in this discipline, as in other areas of clinical practice.

It is biologically implausible that there is any innate difference in ability between peoples of different cultural backgrounds. Such differences are far more likely to lie in socio-economic and societal factors, including hidden implicit biases in selection and assessment structures. An excellent and extensive work programme[30] recently identified the causes of differential attainment as bias, social class, deprivation, anti-immigrant mentality and geo-political disadvantage. Among the vectors for these challenges were assessment structures.

Recommendation 19: The impact of gender, ethnicity and educational background on exam performance in the RCoA exams be explored through a research study, with findings incorporated into the ongoing assessment design process.

11.3 Validity research

The meaning of 'validity' in assessment is complex[31], but here I will focus on three relatively simple interpretations of concurrent, predictive and construct validity.

11.3.1 Concurrent validity

How well do the assessments match the experiences and judgements of trainers and educational supervisors? I have posed this question in various forms to respondents in various categories during this Review, and the themes that have emerged are that, if the assessments are viewed as screening tests in medical terms, they are sensitive but not specific. In other words, those who pass are generally viewed as having deserved to pass, with a few exceptions, but a larger number of those who failed were deemed to have been good performers in the workplace, in whose clinical work their trainers had confidence. These are false negatives, which have significant financial and emotional costs to the individual. But false negatives have medical and societal costs too, especially if some candidates leave the profession as a result of being rated as false negatives.

Assessment outcomes could be compared with Structured Learning Events (WPBAs), but these are not perfect instruments, as they currently are largely formative and may suffer from the phenomenon of 'failure to fail'[32]. Rather a research study should be conducted to compare the confidential estimation of trainers and supervisors of candidates' capability in the workplace with the performance of those candidates in the various aspects of the exams.

This approach would greatly empower the process of standard setting, which at the moment takes place largely independently of evidence on actual performance in the clinical workplace. Use of the

Hofstee standard setting method, as recommended in this Review, would also bring standard setting processes closer to workplace estimates of ability.

Recommendation 20: A concurrent validity study be conducted to compare performance in the workplace as estimated by trainers and supervisors with performance in the RCoA assessments, with findings incorporated into the ongoing assessment design process.

11.3.2 Predictive validity

Predictive validity relates to how well assessment predict later events, especially those in the workplace. It would be possible to compare, for instance, performance in RCoA assessments with ARCP outcomes and Fitness to Practice outcomes through the UK Medical Education Database[33], career outcomes such as time to appointment to a consultant post and even, what is probably the gold standard, outcomes for patients, although this data is difficult to collect (see for example Norcini et al., 2022[34]).

Recommendation 21: A predictive validity study be conducted to compare performance in the workplace as estimated by trainers and supervisors with performance in the RCoA assessments, with findings incorporated into the ongoing assessment design process.

11.3.3 Construct validity

One way to explore construct validity is to choose an assessment and have it undertaken not just by the immediate candidates but by individuals at a number of different stages of their careers (including trainers and supervisors, and, possibly, even RCoA examiners not directly involved in producing that part of the assessment). There should be a positive relationship between stage of seniority and performance in the exam.

Interestingly, this concept was spontaneously proposed by several respondent who volunteered to sit the assessments as candidates again. This would help address one of the key problems of face-to-face assessment, which is that examiners necessarily have advance knowledge and preparation time for the questions, while candidates experience them unexpectedly in circumstances which may not lend themselves to knowledge retrieval. Several examiners described their advance study for questions arising in the SOE to me, and this is a mark of their conscientiousness. Yet candidates are then presented with the questions without advance knowledge and without the context of clinical settings which may aid their knowledge retrieval in a real-world setting.

Recruitment of examiners and senior colleagues may prove an interesting challenge, even where the outcome is confidential, but it is not my intention to design any of these research projects in detail in this Review.

Recommendation 22: A construct validity study be conducted to compare performance of candidates with colleagues at different stages of their professional careers, with findings incorporated into the ongoing assessment design process.

12. Equality, diversity and inclusion

Concerning EDI issues, issues were recounted to me from both the examiner and examinee sides, particularly with regard to reasonable adjustments.

Within the RCoA and Faculties, the mechanism for establishing policy on EDI seemed unclear. Decisions concerning reasonable adjustments seemed not to draw on a clear or codified policy

within which particular individual needs could be explored. Several individuals, including the Head of Examinations, seemed to be involved in making these decisions and engaging in correspondence on these matters. However, accordance with the relevant Equality legislation is something which benefits from dedicated experience.

Section 149 of the Equality Act 2010 provides that:

"(1) A public authority must, in the exercise of its functions, have due regard to the need to—

(a) eliminate discrimination, harassment, victimisation and any other conduct that is prohibited by or under this Act;

(b) advance equality of opportunity between persons who share a relevant protected characteristic and persons who do not share it;

(c) foster good relations between persons who share a relevant protected characteristic and persons who do not share it.

(2) A person who is not a public authority but who exercises public functions must, in the exercise of those functions, have due regard to the matters mentioned in subsection (1).

(3) Having due regard to the need to advance equality of opportunity between persons who share a relevant protected characteristic and persons who do not share it involves having due regard, in particular, to the need to—

(a) remove or minimise disadvantages suffered by persons who share a relevant protected characteristic that are connected to that characteristic;

(b) take steps to meet the needs of persons who share a relevant protected characteristic that are different from the needs of persons who do not share it;

(c) encourage persons who share a relevant protected characteristic to participate in public life or in any other activity in which participation by such persons is disproportionately low.

(4) The steps involved in meeting the needs of disabled persons that are different from the needs of persons who are not disabled include, in particular, steps to take account of disabled persons' disabilities. ...

(7) The relevant protected characteristics are: age; disability; gender reassignment; pregnancy and maternity; race; religion or belief; sex; sexual orientation"

This is known as the 'public sector equality duty' (PSED).

The Equality and Human Rights Commission (EHRC) has produced technical guidance in respect of the PSED. The aim of this guidance (2.10:17) is set out as follows.

> The broad aim of the general equality duty is to integrate consideration of the advancement of equality into the day-to-day business of all bodies subject to the duty. The general equality duty is intended to accelerate progress towards equality for all, by placing a responsibility on bodies subject to the duty to consider how they can work to tackle systemic discrimination and disadvantage affecting people with particular protected characteristics There are three aspects to the duty – elimination of discrimination, promotion of equality of opportunity and fostering of good relations.

The term 'due regard' in the Equality Act does not mean that every candidate should have identical access to the exam process. There should be a process of weighing and assessing practicalities and consequences in arriving at reasonable adjustments. The best way to demonstrate that this has

happened is via an Equality Impact Assessment (EIA) which, while not a requirement of the Equality Act, would serve as evidence that 'due regard' had been paid to the Act.

While I understand that major changes requiring GMC review were the subject of EIAs, I believe that significant changes short of those requiring GMC review should also be subject to an EIA. Ongoing assessment changes in contemplation would also benefit from extensive candidate input.

I therefore recommend that an Equality, Diversity and Inclusion Lead be appointed within the College, to establish consistent policy across all assessments. This need not be a salaried post. In my experience, within organisations, there are often individuals with a special interest in EDI, and it may be that a volunteer can be found among the examiners or elected members of the College and Faculties.

Recommendation 23: That an Equality, Diversity and Inclusion Lead be appointed to lead on EDI issues across the College and Faculties.

Recommendation 24: That Equality Impact Assessments be conducted with regard to significant planned changes in the RCoA assessment processes.

13.Examiner culture

While I found no evidence of systematic bias against candidates on the basis of protected characteristics, there was clear evidence that in the past the social behaviour of some examiners had posed challenges for others in both digital and live settings. These challenges, while not rising to the level of outright sexism or prejudice, had caused discomfort and unhappiness in the examiner body, and this discomfort has not been fully appreciated and understood by some long-serving examiners. While, by the account of many individuals, the situation has improved markedly, residual sexism and bias remain something to be guarded against, and this should be addressed through appropriate training. In terms of the College values, including examiner culture in this review indicates openness and responsiveness, as do the development of Codes of Conduct and the behavioural changes that took place as a result of these. Plainly, over recent years the College has become more diverse and inclusive, and this in itself will help remedy some of the problems that have arisen. Efforts to this end should continue and be enhanced, to ensure that the College is, indeed, caring, supportive, just and fair in its endeavours.

Recommendation 25: That formal assessment communication channels in the College are used solely for assessment purposes.

Recommendation 26: that examiner training explicitly reinforces the requirement for professional communication during examinations and the risks that arise from unprofessional behaviour, and that the Codes of Conduct form a key part of this training.

Recommendation 27: that the mentoring of new examiners continues, and that examiner training includes discussion of the importance of avoiding a hierarchical approach and language in exam discussions.

Recommendation 28: that efforts to increase the diversity of the examiner pool, especially in senior positions, continue and are extended.

Summary list of Recommendations

Recommendation 1: If a marked discrepancy from previous results is observed without an identifiable causative factor, the Hofstee compromise method be employed as the primary standard setting approach.

Recommendation 2: The Head of Examinations or colleagues to which they delegate responsibility become a full member of all relevant assessment committees, including Faculty bodies, in a decision-making role on a par with other committee members.

Recommendation 3: Candidate and doctors-in-training representation in an advisory capacity on all RCoA committees engaged in assessment be assured, and these representatives have a key role in supporting communication with candidates.

Recommendation 4: Clear lines of governance and accountability should exist across all the College and Faculty assessment processes.

Recommendation 5: Identification of best assessment practice on issues such as standard setting, item writing, and item banking should be followed by the application of these practices consistently across all the assessments run under the aegis of the RCoA.

Recommendation 6: A clear statement of the intended purpose of the assessment be drawn up for all the assessments, and that the results of benchmarking by doctors in practice at various levels be used to inform the standard setting procedures.

Recommendation 7: The written components of all RCoA assessments are based on Single Best Answer MCQs, rather than Multiple True False or Constructed Response Questions, as soon as possible.

Recommendation 8: The 'knowledge' stations and components currently present in the OSCE be moved to the written papers.

Recommendation 9: The examiner guides for the OSCE encourage the exercise of judgement within the context of the marks scheme in interpreting candidate responses when they are clearly on the right lines.

Recommendation 10: (a) The 'knowledge' components of the SOE be moved to the written papers; and (b) the clinical reasoning skills components of the SOE be placed within the format of the OSCE.

Recommendation 11: Commercial assessment platforms are kept under constant review, with a short- to medium-term view to implementing a single platform across all assessments delivered by the RCoA team. This will empower the collection and analysis of both biometric and psychometric data.

Recommendation 12: Common style guides for MCQs and OSCEs be introduced across all assessments delivered by the RCoA team.

Recommendation 13: Common high quality training materials and approaches be employed across all assessments delivered by the RCoA team, and that a culture of continuous reflection and improvement be maintained.

Recommendation 14: Inclusive recruitment of examiners should continue to be promoted by the College, as a personal, professional and societal benefit, and the requirements for becoming eligible to be an examiner are reviewed with a view to inclusion.

Recommendation 15: A unified approach to the use of, and training in, standard setting be employed across all the assessments delivered by the RCoA team.

Recommendation 16: The practice of subtracting a Standard Error of Measurement should cease, and guessing corrections be employed consistently across the RCoA exams, and should cease as soon as MTFs have been phased out in favour of SBAs.

Recommendation 17: The Hofstee Compromise Method be used to standardise set assessments while the recommended changes to the RCoA assessment structures are taking place, in order to ensure there are no inappropriate swings in pass/fail rates resulting from these changes.

Recommendation 18: Professional psychometric capability be added to the College to support the Exam Team and exams' committees.

Recommendation 19: The impact of gender, ethnicity and educational background on exam performance in the RCoA exams be explored through a research study, with findings incorporated into the ongoing assessment design process.

Recommendation 20: A concurrent validity study be conducted to compare performance in the workplace as estimated by trainers and supervisors with performance in the RCoA assessments, with findings incorporated into the ongoing assessment design process.

Recommendation 21: A predictive validity study be conducted to compare performance in the workplace as estimated by trainers and supervisors with performance in the RCoA assessments, with findings incorporated into the ongoing assessment design process.

Recommendation 22: A construct validity study be conducted to compare performance of candidates with colleagues at different stages of their professional careers, with findings incorporated into the ongoing assessment design process.

Recommendation 23: An Equality, Diversity and Inclusion Lead be appointed to lead on EDI issues across the College and Faculties.

Recommendation 24: Equality Impact Assessments be conducted with regard to significant planned changes in the RCoA assessment processes.

Recommendation 25: Formal assessment communication channels in the College are used solely for assessment purposes.

Recommendation 26: Examiner training explicitly reinforces the requirement for professional communication during examinations and the risks that arise from unprofessional behaviour; the Codes of Conduct form a key part of this training.

Recommendation 27: The mentoring of new examiners continues, and examiner training includes discussion of the importance of avoiding a hierarchical approach and language in exam discussions.

Recommendation 28: Efforts to increase the diversity of the examiner pool, especially in senior positions, continue and are extended.

About the author

The author is currently Professor of Medical Education and formerly Head of School at UCLan Medical School. He is a GMC Associate and has carried out a number of projects commissioned by the GMC (with regard to the Professional and Linguistic Assessment Board tests for International Medical Graduates), by Health Education England (with reference to the Royal College of General Practitioners assessment structure), by the Department of Health (as academic partner reviewing revalidation for doctors), amongst others. Currently he serves on the Expert Reference group for the GMC's Applied Knowledge Test, part of the proposed national Medical Licensing Assessment, and is Psychometric Advisor to the Recruitment Development Group of the UK Foundation Programme Office. Previously he was a Board Member of the UK Clinical Aptitude Test, and Editor-in-Chief of *Medical Education*, the leading journal in the field. He has published widely on assessment in health care settings. In 2022, he was awarded the Gold Medal of the Association for the Study of Medical Education for his services to the field.

References

[1] Frankland, P.W., Josselyn, S.A. and Köhler, S., 2019. The neurobiological foundation of memory retrieval. *Nature neuroscience*, *22*(10), pp.1576-1585.

[2] Case S, Swanson D (2002) Constructing written test questions for the basic and clinical sciences. National Board of Medical Examiners, Philadelphia.

[3] Downing, S.M., 1992. True-false, alternate-choice, and multiple-choice items. *Educational Measurement: Issues and Practice*.

[4] Brabec, J.A., Pan, S.C., Bjork, E.L. and Bjork, R.A., 2021. True-False Testing on Trial: Guilty as Charged or Falsely Accused? *Educational Psychology Review*, *33*(2), pp.667-692.

[5] Schmidt, D., Raupach, T., Wiegand, A., Herrmann, M. and Kanzow, P., 2021. Relation between examinees' true knowledge and examination scores: Systematic review and exemplary calculations on Multiple-True-False items. *Educational Research Review*, p.100409.

[6] Krathwohl, D.R., 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice*, *41*(4), pp.212-218.

[7] Ikah, D.S., Finn, G.M., Swamy, M., White, P.M. and McLachlan, J.C., 2015. Clinical vignettes improve performance in anatomy practical assessment. *Anatomical Sciences Education*, *8*(3), pp.221-229.

[8] Schuwirth, L.W.T., Van der Vleuten, C.P.M. and Donkers, H.H.L.M., 1996. A closer look at cueing effects in multiple-choice questions. *Medical Education*, *30*(1), pp.44-49.

[9] Schuwirth, L.W. and Van Der Vleuten, C.P., 2004. Different written assessment methods: what can be said about their strengths and weaknesses?. *Medical education*, *38*(9), pp.974-979.

[10] Malau-Aduli, B.S., 2011. Exploring the experiences and coping strategies of international medical students. *BMC medical education*, *11*(1), pp.1-12.

[11] Sam, A.H., Westacott, R., Gurnell, M., Wilson, R., Meeran, K. and Brown, C., 2019. Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. *BMJ open*, *9*(9), p.e032550.

[12] Lord FM (1959) Tests of the same length do have the same standard error of measurement. Educational and Psychological measurement 19:233-239.

[13] Boursicot, K., Kemp, S., Wilkinson, T., Findyartini, A., Canning, C., Cilliers, F. and Fuller, R., 2021. Performance assessment: Consensus statement and recommendations from the 2020 Ottawa Conference. *Medical Teacher*, *43*(1), pp.58-67.

[14] Homer, M., 2022. Pass/fail decisions and standards: the impact of differential examiner stringency on OSCE outcomes. *Advances in Health Sciences Education*, pp.1-17.

[15] nbme.org/item-writing-guide

[16] https://www.aomrc.org.uk/reports-guidance/standard-setting-framework-postgrad-exams-1015/

[17] Monteiro, S., Sullivan, G.M. and Chan, T.M., 2019. Generalizability theory made simple (r): an introductory primer to G-studies. *Journal of graduate medical education*, *11*(4), pp.365-370.

[18] Van der Linden, W.J. and Hambleton, R.K., 1997. Handbook of item response theory. *Taylor & Francis Group.*

[19] Jahshan, A., Aoun, M., Dekhou, A. and Folbe, A., 2021. The underrepresentation of women and ethnic minorities in anesthesiology. *Journal of the National Medical Association*.

[20] Bosco, L., Lorello, G.R., Flexman, A.M. and Hastie, M.J., 2020. Women in anaesthesia: a scoping review. *British journal of anaesthesia*, *124*(3), pp.e134-e147.

[21] Wong, C.A., Moonesinghe, S.R., Boer, C., Hemmings, H.C. and Hunter, J.M., 2020. Women in anaesthesia, a special issue of the British Journal of Anaesthesia. *British journal of anaesthesia*, *124*(3), pp.e40-e43.

[22] Bowhay, A.R. and Watmough, S.D., 2009. An evaluation of the performance in the UK Royal College of Anaesthetists primary examination by UK medical school and gender. *BMC Medical Education*, *9*(1), pp.1-8.

[2323] Tsugawa, Y., Jena, A.B., Figueroa, J.F., Orav, E.J., Blumenthal, D.M. and Jha, A.K., 2017. Comparison of hospital mortality and readmission rates for Medicare patients treated by male vs female physicians. JAMA internal medicine, 177(2), pp.206-213.

[24] Wallis, C.J., Ravi, B., Coburn, N., Nam, R.K., Detsky, A.S. and Satkunasivam, R., 2017. Comparison of postoperative outcomes among patients treated by male and female surgeons: a population based matched cohort study. BMJ, 359.

[25] Wallis, C.J., Ravi, B., Coburn, N., Nam, R.K., Detsky, A.S. and Satkunasivam, R., 2017. Comparison of postoperative outcomes among patients treated by male and female surgeons: a population based matched cohort study. BMJ, 359.

[26] Greenwood, B.N., Carnahan, S. and Huang, L., 2018. Patient–physician gender concordance and increased mortality among female heart attack patients. Proceedings of the National Academy of Sciences, 115(34), pp.8569-8574.

[27] Jones, L.K., Jennings, B.M., Higgins, M.K. and De Waal, F.B., 2018. Ethological observations of social behavior in the operating room. *Proceedings of the National Academy of Sciences*, *115*(29), pp.7575-7580.

[28] O'Neill, T.R., Wang, T. and Newton, W.P., 2022. The American Board of Family Medicine's 8 years of experience with differential item functioning. *The Journal of the American Board of Family Medicine*, *35*(1), pp.18-25.

[29] Watmough, S. and Bowhay, A., 2011. An evaluation of the impact of country of primary medical qualification on performance in the UK Royal College of Anaesthetists' examinations. *Medical teacher*, *33*(11), pp.938-940.

[30] Chakravorty, I., Daga, S., Sharma, S., Chakravorty, S., Fischer, M. and Mehta, R., 2021. Bridging the Gap 2021-Report. *Sushruta Journal of Health Policy & Opinion*, pp.1-107.

[31] Downing, S.M., 2003. Validity: on the meaningful interpretation of assessment data. *Medical education*, *37*(9), pp.830-837.

[32] Cleland, J.A., Knight, L.V., Rees, C.E., Tracey, S. and Bond, C.M., 2008. Is it me or is it them? Factors that influence the passing of underperforming students. *Medical education*, *42*(8), pp.800-809.

[33] www.ukmed.ac.uk

[34] Norcini JJ, Weng W, Boulet J, et al. (2022) Associations between initial American Board of Internal Medicine certification and maintenance of certification status of attending physicians and in-hospital mortality of patients with acute myocardial infarction or congestive heart failure: a retrospective cohort study of hospitalisations in Pennsylvania, USA. BMJ Open 2022;12:e055558. doi:10.1136/ bmjopen-2021-055558