Review of the Assessment Procedures of the Royal College of Emergency Medicine

Professor John C. McLachlan

February 2023

Contents

Executive summary

In March 2022, a sitting of the Royal College of Emergency Medicine Fellowship Single Best Answer paper (FRCEM SBA) was undertaken by candidates, and results were released on April 22nd, 2022. Following queries from candidates who were reviewing their feedback, it emerged that due to an error in handling results, fifty candidates had been informed that they had passed when, in fact, they had failed.

As a result, the College determined to institute major changes in the assessment processes and to arrange an independent external review with regard to both existing practice and changes currently being made, to benefit from advice on best practice in the field. The author was commissioned to carry out this review. The agreed methodology was mainly confidential semi-structured interviews and document review. The College provided contact details for key individuals, and an open invitation was extended to key stakeholders, including individual candidates, trainees and candidate group representatives such as the Emergency Medicine Trainee's Association (EMTA) and the Forum for Emergency Medicine Specialty and Specialist Doctors (EMSAS) to participate in the review. Fortnightly meetings were held with the CEO and Director of Education throughout the period of the review. These were to resolve queries and update on progress. Many of my developing recommendations brought forward at these meetings either paralleled changes already under way at the College or were incorporated into the change process. As a result, a number of my recommendations are already being acted upon.

Concerns about poor communication, loss of trust in the College, and a feeling that the College did not care about them, were expressed to me by trainees. But in fact, I found that in the College, from the President down, deep regret for the mistake, frustration and concern that it had been made and sympathy for all affected candidates were universally expressed. There was recognition that things had to change radically, and many major changes, in both processes and roles, had already commenced when I began my review.

In my opinion, after reviewing the evidence, the error was due to two major stressing factors, the increasing candidate numbers and the pandemic, coupled with organisational weaknesses in the exam processes of the kind known in safety studies as latent errors. These together resulted in an error occurring, then being missed in internal checking processes and finally being communicated to candidates. Blame for particular individuals is therefore not part of the outcome of this review. While individual errors were made, I believe that these are comparable to the kinds of human error that occur in the clinical setting, when individuals are under particular stress.

I hope this report might promote understanding on the part of stakeholders, including the College officers and staff, on one hand, and candidates for exams on the other, of the challenges that each face in navigating these difficult issues in the time of the pandemic and at a time of massive change.

As in all assessment in healthcare, there is a problem in the College of differential attainment by candidates with protected characteristics, in both knowledge and skills tests. This is an intractable challenge with no obvious opportunities for speedy resolution, but some remarks on this topic are made in the appropriate section. Some comments were received about the handling of reasonable adjustments, and I have explored this and Equality, Diversity and Inclusion (EDI) issues, including equality impact assessments.

Assessment formats are appropriate. Many changes in curriculum and exam format have taken place recently, and as a result a degree of stability would be welcomed by candidates. I do not propose changes to delivery mechanisms such as the online format and face-to-face Objective Structured Clinical Examination (OSCEs) but will comment on possible challenges to security and to exam malpractice issues.

With regard to assessment delivery platforms, I recommend that the College continues its efforts to simplify the software systems used, and a continuous review of the software market. The information flow through the system is still complex and requires manual handling at various points, and I recommend simplification of this where possible.

Major changes have been made to the internal organisation, to quality auditing and to process mapping. I am confident that these changes will bring long term benefits. In the short term, there are risks in implementing changes even if the long-term benefit is clear. There is still a window of risk for the College. However, officers and staff seemed fully aware of this as part of the process.

I devote some consideration to the purposes and standards of the assessments and make recommendations about exploring the predictive validity of the exams, and how well they match actual performance in the workplace.

It is clear to me that many of the problems, both practical and philosophical, faced by RCEM, are common to medical Royal Colleges in general. I therefore make some recommendations aimed at addressing information sharing and development in areas such as exam materials and processes, along the lines of the Medical Schools Council Assessment Alliance (MSCAA).

The Royal College of Emergency Medicine (RCEM) is not unique, even among Royal Colleges, in suffering from challenges in the delivery of complex assessment processes, particularly during the time of the pandemic. The Academy Trainees Doctors' Group has recently proposed a well-considered set of practices for Royal Colleges in cases where an assessment error has arisen, and these echo many of the themes in this review.

1. Introduction

1.1 Background to the Review

The Royal College of Emergency Medicine was incorporated by Royal Charter in 2008. It is registered with the Charity Commission and the Scottish Charity Commission. Its purpose is to advance education and research in Emergency Medicine. The College is responsible for setting standards of training and administering examinations in Emergency Medicine for the award of Fellowship and Membership of the College as well as recommending trainees for the completion of their training in Emergency Medicine. The College works to ensure high quality care by setting and monitoring standards of training and education. It provides expert guidance and advice on policy to relevant bodies on matters relating to Emergency Medicine and makes representations to Government and NHS bodies with regard to patient safety and doctor practice.

The RCEM examinations are part of the postgraduate medical training programme in the UK and are approved by the General Medical Council for this purpose. Membership and Fellowship examinations are administered by the College.

The Membership Examinations consist of the MRCEM Primary Examination, the MRCEM Intermediate Examination and the MRCEM Objective Structured Clinical Examination (OSCE).

The MRCEM Primary Exam (known as the FRCEM Primary until July 2021) consists of a 3-hour Single Best Answer (SBA) 'One best of five' multiple choice question paper, with 180 items. The MRCEM Intermediate Examination consists of two SBA papers, each lasting two hours. The MRCEM OSCE consists of 16 8-minute stations, plus two rest stations.

The Fellowship Examinations consist of the FRCEM SBA Examination and the FRCEM OSCE.

The FRCEM SBA Examination consists of two SBA papers, each lasting two hours. The FRCEM OSCE consists of 16 8-minute stations, plus two rest stations.

Since the pandemic, RCEM SBA exams are delivered electronically online, with candidates attending a Pearson Vue Test Centre or, in some circumstances, having the option to take the assessment from a location of their own choice.

In March 2022, a diet of the FRCEM SBA was undertaken by candidates, and results were released on April 22nd, 2022. Following queries from candidates who were reviewing their feedback, it emerged that due to an error in results handling, 50 candidates had been informed that they had passed when they had in fact failed.

As a result, the College decided to arrange an independent external review of their assessment processes, with regard to both existing practice and changes currently being made, to benefit from advice on best practice in the field. The author was commissioned to carry out this Review, and the agreed methodology is summarised in the next section.

1.2 Methodology: how the review was carried out

There were four main approaches to conducting the review.

1.2.1 Interviews

A dedicated e-mail account was created for the purpose of confidential communication. Virtual interviews via Teams or Zoom were carried out with key individuals involved with the RCEM assessment processes, with an initial contact list of key individuals being provided by the College. An open invitation was extended to key stakeholders, including individual candidates, trainees and candidate group representatives such as the Emergency Medicine Trainee's Association (EMTA) and the Forum for Emergency Medicine Specialty and Specialist Doctors (EMSAS) to participate in the review.

An open invitation was further 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 were generally scheduled for one hour. Hand-written notes were taken during the meeting, which I expanded 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.

Forty-six direct invitations were issued to individuals and organisations to contribute, and in the end, I spoke with over 60 individuals, including some representing larger groups of individuals such as the Emergency Medicine Training Association and the Forum for Emergency Medicine Speciality and Specialist Doctors.

Fortnightly meetings were arranged with the Chief Executive Officer and the Director of Education. These were originally intended to keep them up to date on the general progress of the review, but in practice proved even more useful than expected, and as a result a number of suggestions and recommendations which I made were integrated into the College's continuous improvement strategy. As a result, while I will still make formal recommendations, a number of these are already in train, and they may indeed echo those which the College is already making.

1.2.2 Document Review

A secure Teams area was created, into which the College could place documentation relating to all areas of the examination process. Many of these documents were placed in this proactively by the College, but in addition there were documents whose existence I identified in the course of discussions, and copies of which I requested. All such requests were met by the College in a timely manner. Documents included meeting minutes, previous reviews, correspondence, surveys and spreadsheets. These were reviewed and key points were used during the interviews and in writing this final report.

1.2.3 Surveys

I was given access to a variety of feedback materials obtained from candidates and College members. Considering the open invitation to contact me and of the feedback material with which I was provided, a separate survey was not circulated to stakeholders.

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.

2. General remarks

In the course of the review, the significant hurt caused to candidates and potential candidates by errors in exam processing became clear, and a lack of confidence in the College was expressed by a number of respondents. Candidates felt that the College showed a lack of demonstrated understanding of the impact that exams can have on candidates' lives in terms of time commitment while working in a challenging environment, family and social disruption, stress, and financial cost.

In particular, there was a feeling from trainees that initially the College did not always fully appreciate the importance of the FRCEM SBA as an exit exam. Individuals who believed they had passed could, for instance, have been offered consultant posts on that basis or planned or even initiated house moves, with or without

family. Having told colleagues and family members that they had passed, and having had a moment of celebration, made it even more difficult to have to retract this outcome. Some candidates expressed an ongoing lack of trust in College outcomes, indicating that even if they have been told they have passed an exam, they may have doubts as to whether this might subsequently prove not to have been the case.

It was noted that the initial response did not include a clear apology (See Section 3.4). However, it was also felt that the subsequent response (detailed in Section 3.4) was perhaps the best it could have been. This degree of understanding may be in part due to the nature of the specialty, with EM doctors being used to decision making in complex, time-limited situations.

Communication with RCEM was frequently perceived as poor. There appear to have been delays in replying to e-mails, with significant backlogs building up, and difficulty in candidates getting through to the College by other means.

However, I saw no evidence that the individuals who make up the College were uncaring. Real understanding, concern and sympathy were widely expressed by College officers and staff for candidates and trainees as a group and with particular respect to the major FRCEM SBA error. In general, significant improvements to the exam processes have already been made to address the concerns, and these are ongoing. I have made a number of further recommendations for change in the subsequent pages of this report.

The actual content and the original scoring of the exams was not affected: the March 2022 error arose from the subsequent reporting of the results. As a consequence, no challenge to patient safety was posed by candidates passing who did not deserve to pass.

Reasons why the error occurred are explored in Section 3.3. I note that the College was under considerable stress, in terms of changes to the assessment structure, the increasing number of candidates and the impact of the pandemic, including the need to switch assessments to online versions at short notice. The flow of information through a variety of platforms was complex and insecure. Human errors did occur against this challenging background.

A major aim of my report is to support reconciliation and increased understanding and trust between the College and the candidates for its exams. It is clear to me that the College indeed aims to work in the best interests of candidates, and one senior member expressed his/her desire that the College should be a haven and support for trainees of all kinds, rather than a challenge. In the current state of health delivery, Emergency Medicine doctors and Advanced Care Practitioners play particularly vital roles in healthcare delivery, and good relationships within the College, and between the College and the candidates, are of inestimable value, not merely to the individuals concerned, but also to the well-being of the population as a whole.

3. Issues with the March 2022 FRCEM SBA

3.1 What normally happens in exam development, delivery and reporting?

It is essential first to describe the normal information flow with regard to the SBA examinations. The question bank is first created in commercial software known as the assessment platform, then the exam is manually extracted to separate documents and sent to Pearson Vue, who deliver the exam to candidates mostly via their network of test centres. A small proportion of candidates sit the assessment from their own location, online via the OnVue platform. Pearson Vue build the test in their test publisher for administration to candidates. After the exam is complete, the results are exported by Pearson Vue to the psychometrician for manipulation (re-structuring) into a file format suitable for candidate responses to be reloaded to the assessment platform, where they are scored. That scored file is exported to be used in an adjudication meeting, at which items are identified for removal. This takes place in the assessment platform. Then the file is re-exported to be uploaded into the iMIS database, where candidates can log in to see their results. These results contain generic feedback, from which it is possible for candidates to calculate their scores. A different platform, Practique from Fry IT, was used for delivery of the OSCEs.

3.2 What happened after the March 2022 FRCEM SBA?

On 13th April 2022, candidates who had undertaken the March 15th FRCEM 2022 SBA MCQ exam were given their assessment results. Subsequently candidates got in touch with RCEM to query the addition of marks on

their feedback. On investigation, it proved that there had been an error in the processing of the results, and fifty candidates had been informed that they had passed the assessment when, in fact, the outcome should have been that they had failed.

3.3 How did it happen?

3.3.1 Specific factors

I heard several slightly differing accounts of exactly what happened, each reflecting the knowledge of a number of different individuals. However, the following account represents the most commonly held view on the events, though it is not quite a consensus.

After the exam had been completed, the paper was reviewed in terms of the performance of each question within it. This is a standard practice and aims to identify questions which have not performed well. For instance, almost every candidate may have got a particular question wrong, apart from random guessing. For such a highly qualified group of candidates, if everyone gets the question wrong, the problem is probably with the question, not the candidates. Previously with the RCEM system, problems could have arisen with an image: a radiograph which had been quite clear on the original test material may not have transferred well through all the file format changes which were then required, so by the time it got to candidates in a Pearson Vue centre, it was not possible to draw the correct conclusions from it. Images are now handled in a better way.

Removing these very poorly performing items creates what is effectively a new (and in some ways better) exam paper. I will comment further on this, in a more technical manner, in Section 6.5.2.

In this particular exam, eleven items out of a total of 180 were removed in this redaction process.

Normally, these decisions are automatically acted on by the assessment software, which recalculates candidates' scores and their new 'pass/fail' outcomes.

However, for some reason, which remains unclear (it is suggested that the software system, the assessment platform, failed to update in response to items being removed), this automatic procedure did not happen, so while the pass mark was recalculated, the original scores remained unchanged. Technically, since this represents something going wrong within the software, this would represent a failure rather than an error[1].

Here is a simplified illustration of how this affects the outcomes. Let us imagine that the original paper has 180 items, and the pass mark is 60% (i.e. candidates must get 108 questions correct out of 180). A candidate who scores 105 has, therefore, failed: they only scored approximately 58% on the exam. If 11 questions are removed for good reason, the exam now has 169 items rather than 180, and the 60% pass mark now represents a requirement to get approximately 101 questions correct.

A candidate who got 58% correct in the original paper will now, on average, score 58% of 169 for the corrected paper, and therefore get about 99 questions correct. They will still have failed.

But if by some error their original uncorrected score of 105 was transferred to the new exam, where the 60% 'pass' requirement is for a score of 101, it would look as if they had passed. This is essentially what happened to 50 candidates, who were then told they had passed, when, in fact, they had failed.

Of course, as one might expect, the complete story is a little more complicated than that, and some of these complexities are explored in Section 6.5.2, but this is the essence of the matter. If the original scores before redaction are transferred to a paper whose pass mark was affected by removing 11 items, some candidates who ought to have failed will be reported as having passed. In essence this is what happened to that most unfortunate fifty.

Exactly how the software failed in this unusual way remains unclear. The software company, which provided the assessment platform, had previously recommended that the software be updated, but this had not happened. There is a possibility that a manual handling error occurred instead of a software error, but there is no way retrospectively to establish if this happened, or if it was indeed a software failure.

Normally, the output would be checked by an experienced person. However, the individual administrator doing the first-line check was new to the task and had been in post for less than three weeks. In principle, a second check should also have been made by another member of the administrative staff, but if this took place, the error was not identified here either. At this point, the absence of clear written flow charts for

administrative processes, especially when new staff are involved, and when systems had changed as a result of the pandemic, is a likely contributing factor.

It should be noted that a software failure of the kind described above is very unusual, and it would not have been anyone's normal expectation that this would be a risk. This may have led to overconfidence in the computer output.

In some other organisations, the checking would be done by a psychometrician, who would be more likely to spot an error through familiarity with the rationale behind the processes that were being carried out.

The inaccurate results were then processed through the system and released to the candidates. Note that the process as described is asymmetric: if the pass mark used was too high, then candidates might be incorrectly told they had passed, but no candidates would be told they had failed when they had in fact passed.

3.3.2 General factors which led to the errors

3.3.2.1 Complexity of communications systems, particularly data flow

The data flow in RCEM is complex, with different platforms being involved at different points. This results in increased manual handling, and at each manual handling step, the possibility of human error is introduced. At some stages, for instance, Excel spreadsheets have to be manually re-configured before being re-delivered to one of the software platforms.

3.3.2.2 Communication between different work areas of the College

The exam management process was described to me as rather isolated, indeed, operated in something of a silo from other parts of the College. There was a perception that they were the technical experts in assessment. Communications with other parts of the College which were in fact relevant to exams, such as candidate communication, IT systems and information flow, were insufficient.

3.3.2.3 Resourcing for the exam process

Since the workload for the College, particularly the exams processes, had increased substantially, it is possible that resources were not fully adequate for the tasks required. A business case for provision of further resources had previously been put forward, but the consensus I heard was that it had not been well set out by the exams management and, as a result, it had not been fully funded. Here perhaps was an example of how the tendency of the exams team to work in isolation had rebounded to their disadvantage.

The marked expansion of the College, particularly with regard to assessments, was undoubtedly a factor in straining capabilities. Of course, it was also a marker of the College's success, particularity internationally. But given RCEM's relatively recent inception as a College, workflows which worked well at small scale may not have worked well at larger scale. An example would be the absence of clearly written and adhered to process maps and protocols. At small scale, an organisation where everyone is familiar with their job and knows everyone else personally, and jobs are generally manageable, may not require fully written down procedures. At a larger scale, these become essential. But the point at which this happens may be discovered only by something going wrong, as is perhaps the case here.

The exam calendar was very full as a result of this expansion, and this cut into the time and capacity for checking results for each individual exam.

3.3.2.4 The impact of the pandemic

The ongoing pandemic and the measures taken to alleviate it are likely to have had a significant effect. While, of course, the main burden, as far as the College was concerned, fell on EM doctors in clinical practice, all staff would have been affected in many ways, including workloads, illness, bereavement, isolation, and so on. The College as a whole had had to transition to online and test-centre delivery of the exams at very short notice, at a time when the national resources to support distance assessment were themselves strained by the move of so many providers into the digital environment.

3.3.2.5 Neglect of warning signs

Pressure of work, the isolationism of the exams team and the move to online working may have concealed the significance of a number of smaller scale errors that had previously occurred in the exams process (described to me as 'foreshocks') before the 'mainshock' of the March events. From Spring 2018 onwards there were a variety of less significant errors, which the exams team generically tended to blame on 'IT problems', obscuring

the significant quality audit processes which were in fact failing. The software company, which provided the assessment platform, had in fact previously recommended updates to the locally hosted version of the software to a more recent cloud-based system, but this had not been acted upon.

3.3.2.6 The number of items removed

While issuing the wrong results to candidates is perhaps the worst thing that can happen with regard to a high-stakes professional exam even if only one candidate were affected, the scale of the error is important as well as the severity. In this particular case, fifty candidates were affected. One reason this took place on such a scale was that a significant proportion of items (11/180) were removed from the exam. This would be unusual in a high-stakes setting, with removal of one or two items, or even none, being more common. I will return to this issue in Section 6.5.2.

3.3.2.7 Culture and Communications

The view was expressed to me that sometimes the College may have taken the candidates rather for granted, since the College is the sole provider of the required qualifications. I was told that there was generally a large backlog of unopened e-mails, amounting to over two thousand on occasion, in the exams' inbox. This may have delayed the timely discovery and response to the mistake since it was first noted by candidates rather than the College.

The College subsequently gained a degree of sympathy and understanding from trainees and candidates, once an apology had been issued, and steps to alleviate the problem had been taken. It was suggested that perhaps EM doctors are familiar with high stress situations, with a risk of errors arising under stressful situations.

3.4 What actions were taken in specific response to this error?

Existence of a problem was first identified by a candidate who could not make the results on their feedback tally with the overall total for the exam. An e-mail was sent to the College on 25th April and read on 29th April due to a backlog in the e-mail system. The error then emerged on checking of the results. From now on, events progressed more rapidly. The CEO, President and the Chair of the Governance Committee were informed, and a crisis management team was created, which met daily and sometimes twice daily. The GMC and Charity Commission were informed. With the guidance of the GMC, it was determined that the true results should stand (as opposed to modifying the pass mark in any way) and that consequently candidates who had previously been informed that they had passed would now be told that they had failed.

In terms of communication with candidates, to the frustration of College senior members, the issuing of an immediate apology was delayed for insurance reasons.

Psychological support via counsellors was offered to the affected candidates, and very senior members of the College from the President, CEO and Directors, down made themselves available for discussions with candidates. The College established a special diet in July for those who had been affected and provided additional training including four revision webinars (which I understand will be made available to all candidates). Forty of the fifty affected candidates made use of this special diet. Fees and expenses for the exam were refunded, and the failed exam was discounted from the total number of exams taken.

The President, CEO and other senior staff spoke with trainee organisations and various colleagues, such as Deans, Heads of Schools, and others.

3.5 What has been done internally in response to these challenges?

A major programme of quality audit and improvement has been embarked upon by the College with particular response to these events. During the time of my review, process maps were being developed for all key activities. The exam calendar has been revised so there are fewer separate OSCE exams (although the number of places has increased) to avoid the calendar being so fully occupied and to give more time for checking results before they are issued. A number of personnel changes have been made, and new staff have been brought in.

Software use has been reviewed, and while the same platforms are in use, updates have been made to the version of the assessment platform used, and College staff are working closely with the assessment platform provider to adapt the software to better meet the needs of the College.

There has been a programme of continuous improvement in customer service, resulting in a significant decrease in the number of unopened e-mails. Much more results checking now takes place. However, this has also extended the period before the results can be reported to candidates, and I heard some concerns from candidates about these delays. Hopefully, once the new processes have been fully embedded, this time can be shortened again, so candidates have their results in a timely manner.

I heard reports of openness and frankness on the part of senior staff and generally of positive responses to the changes. Time of change is also stressful, however, and some more junior staff reflected on the challenges they faced with restructuring, and new roles and procedures. Many of the team are still relatively new to the process, and anxieties exist around roles and remits.

Openness and communication between different parts of the College is now appreciated and encouraged. The isolation of the exam team previously described is no longer the case.

3.6 General conclusions and further recommendations: has enough been done?

The College was faced with two major stressors of their systems. These were the expansion in the number of candidates and the switch to online assessment consequent on the pandemic. There were also systemic weaknesses in the assessment processes. These were the complex information flow through the assessment processes, the siloed nature of the Exams Team and the lack of clearly articulated quality assurance process maps, particularly with regard to checking. These were *latent errors*[2], in the terminology of medical error classification.

The impact of Covid, and the required switch to online assessment were unpredictable events and also synergistic with the increasing numbers of candidates. At this point, the intrinsic weaknesses in the system were severely tested and, in consequence, failed. There were undoubtedly individual human errors at the failure points, but I view the sequence of events primarily as a systems failure than as a failure by individuals. Nonetheless, the consequences for individual candidates were severe.

The decision to allow the corrected results to stand was the correct one under the circumstances. Otherwise, concerns may have been raised as to patient safety.

A two day delay was present in issuing the first statement of apologies. While this was in response to the position asserted by the College insurers, it was inappropriate, and the College should have taken more determined action. Once this barrier had been overcome, the College's actions in attempting to alleviate the consequences of the error, as summarised in Section 3.4, were appropriate and were, in general, well received by candidates.

A crisis can sometimes bring positive opportunities, and the major programme of process and management change embarked upon is, I believe, likely to lead to major long-term improvements in the exam processes of the College, and in many ways, of other processes too, such as communication with candidates. Confidence in the senior management of the College, particularly in the light of role and personnel changes, was expressed by many respondents, and I accord with this confidence: the senior management team seemed to me to be capable and, certainly now, well tested by experience.

However, times of major change are stressful, and until all the new processes and team members are bedded in, I believe there remains a window of some vulnerability in which it is not impossible that an error might occur. It may be necessary to go through a complete annual cycle of the exam processes before familiarity begins to bring security to the processes. I heard of a remaining degree of uncertainty about lower level job descriptions, and there still seemed to be a degree of overlap in some roles at the time of interviewing, although this may have been addressed since then. As is inevitable after a major error, anxieties on the part of College staff at all levels about possible future errors remains high. This in itself is a source of stress, and the College should remain cognisant of staff well-being in these circumstances. Open dialogue without recrimination is the best way to address challenges arising under these circumstances, and I heard a number of positive comments along these lines about existing practices.

The information flow through the system remains complex, and I heard of challenges still arising with the software platforms, particularly the assessment platform, Practique and iMIS. Important changes have been made, particularly to the assessment platform, and I will return to these platforms in section 4.1.

Error reporting and analysis and information from appeals, complaints and surveys would benefit from being integrated into a consistent error management structure and action planning, and I will return to this in the next section.

3.7 General considerations on errors

There is an extensive literature on both medical errors and errors in management in general. I considered several different models of how errors occur, including the Rasmussen and Svedung AcciMap technique which is used to graphically represent system-wide failures, decisions and actions involved in accidents[3]. However, in the end, I preferred the familiar 'Swiss Cheese' model promoted by Reason and Weatherall[4], which envisages that there are a number of barriers to an error taking place, but a chance alignment of holes in these barriers results in the error actually eventuating, in a stochastic rather than a deterministic manner.

In this situation, major stressors were the rapid expansion of the College, the rapid pace of change of the assessment processes and the complexity of information flow. The defences were the proper functioning of the software and the human checking processes. The cumulative failure was the result of a failure and an error in these factors.

There is a human temptation to think of errors as deterministic and having only one cause. However, it is more accurate to think of an error as having a number of causes, and this has the benefit of mitigating blame for any one human being at a single decision-making point. Avoidance of premature allocation of blame has positive consequences for good error-management culture[5]. If individuals are able to identify and report the occurrence of errors readily, then early mitigation can reduce the severity of the outcomes and also lead to procedural changes which make future errors less likely.

This is already recognised in the College, as my discussions made clear. Nonetheless, I will go slightly further and suggest that there is a formal process in which reporting an error is welcomed and valued and that this extends to both complaints and appeals from candidates. Consideration should also be given how errors are brought forward, sometimes by quite junior staff. In flying circles, this may be referred to as Crew Resource Management[6]: a healthcare equivalent is SBAR (Situation, Background, Assessment and Recommendation) communication training[7].

Errors, such as those that occurred here, can also be viewed as having positive consequences in identifying potential latent errors and promoting their correction. This is certainly the case in this particular instance, which has brought about marked improvements in RCEM procedures and practices. Admittedly, this is little consolation for the affected candidates, but it is likely that future candidates will benefit from the resulting changes that have been made.

There are other sources of information which can help bring latent errors to light. These include appeals, complaints and surveys. I could not identify a clear process by which these could be analysed over time for emergent themes. For appeals, perhaps development of a standard form which captures the essential information in a structured way would be helpful and would enable a consistent view of how the appeal has progressed. Where an appeal has been successful, this may suggest process improvements that could be made.

Recommendation 1: that a structured appeal process be introduced, so that all appeals and their outcomes could be analysed with a view to identifying procedures in need of improvement.

Recommendation 2: that an integrated system of error management be introduced, encompassing both internally reported errors and externally reported challenges, such as complaints, appeals and surveys, and that communication methodologies such as SBAR be promoted positively to staff and officers to enable the processes of reporting mistakes and challenges.

4. Internal Infrastructure, including IT systems and technology

4.1 Technology platforms

A number of different platforms are involved in transmitting results to candidates. Primarily these are currently the assessment platform for SBAs, which then interface with Pearson Vue's test builder, and Fry IT's Practique used for OSCEs. Output results are delivered to candidates via iMIS.

It may be preferable to have a single platform at least for SBAs and OSCEs, and a number of commercial platforms currently offer this possibility. However, I understand that a recent review of these platforms has taken place and that no universally satisfactory candidate was identified. the assessment platform is currently being upgraded, with active discussion with the company. I understand that a number of improvements will be made to streamline the information flow during this process, such as making the incorporation of images ('assets') more straightforward. The key problem of ensuring that scores update if items are removed will of course be addressed. It will, however, still be necessary to interface with Pearson Vue systems.

While it is recommended that the College continues to scan the horizon for an integrated platform, I do not recommend short-term changes. Many of the College processes are being upgraded in the light of the untoward events of recent times, and this already places a considerable burden on staff. Employment of new software platforms would be a further task which in the short term, may place assessment processes at risk again. A period of stability until the new quality audit and checking processes are in place is essential.

A variety of software platforms are in use elsewhere. Portfolios are delivered via Kaizen (re-named as RISR/Advance during the review, but I will continue to use 'Kaizen' for consistency), and there seemed that there were several computer systems involved in handling refunds. OSANA, Dotdigital and Trello were also mentioned to me in various contexts, particularly the delivery of training materials.

A variety of communications systems with candidates and trainees were also mentioned, including What's App, Twitter, e-mail and others. I was able to discuss the developing communications strategy as part of my review and am satisfied that the problems in communication are acknowledged and are being addressed.

5. Resources and Governance structures

5.1 Were inadequate resources the root cause of the problems?

It is easy to assume that lack of resources was in some way the root cause of the problems that arose. As I indicated in Section 3.7, however, error has many causes. Perhaps it is the case that if more resources had been allocated, this particular problem would not have arisen. But then if the exams team had presented a more convincing business case for more resources, these may have been provided in full rather than in part. And possibly the failure of the exams team to present that convincing case stemmed at least in part from their isolation from the rest of the College processes. As I indicate, I think the major stressors and potential weaknesses in processes were more significant, along with plain unlucky chance. If better induction and support had been available to an inexperienced staff member, if effective further checking of their work had been in place, including a robust sign off process, if the pandemic had not required major changes in exam delivery, if the software had not possibly failed in an unexpected manner, this particular error would perhaps not have happened, though I believe that an error of some kind would have emerged in time.

With regard to the demands on the College, there has already been a well-advised simplification of the exam diet and a reduction of the total assessment burden. New resources have now been allocated to the assessment programme, and, in time, these changes should all work positive benefits.

5.2 Governance and Committee Structures

The general governance structure of the College also did not seem to be at fault. However, one factor that was mentioned to me several times was that there had seemed to be at times a lack of perfect communication at a more personal level between internal constituencies of the College: clinically qualified office holders (often holding very senior and responsible healthcare positions), the employed staff, both senior and junior, the examiners and the lay representatives. The examiners and clinicians saw the exams team at their best, delivering exams in a flexible and helpful way on the ground and were often their advocates. The employed staff were more aware of the isolation of the exams team from feedback into the team and the challenges to

information flowing from this. This culture seems to have changed, and the new leadership of the exams team appears far more open and integrated with the rest of the College. It is important that this continues, and the value of the expertise of each colleague in their special areas is fully recognised, whether this is in clinical matters, in assessment matters such as psychometrics, in exam delivery matters or in communications with candidates.

Trainees and EMSAS are now represented on every RCEM committee, and this is a very welcome step which should help in restoring confidence in the College processes. An annual trainee survey is in place, and it is important that the outcomes from this are treated as part of the error management approach mentioned above in Recommendation 2.

It seems likely that the College will continue with a mixed working pattern for staff, with both home and at-work options available. While this offers many advantages, there is also an issue with the loss of easy and informal communication which presence in a shared physical space empowers. A mixed working model also poses challenges for the induction and enculturation of new staff in particular. The College should recognize the challenges to safe practice this poses, and efforts should be made to ensure that staff have the chance to meet each other in person. Zoom and Teams meetings have a different dynamic to in-person meetings, and a different power differential, and this should be recognised with an appropriate mix of both formats.

Recommendation 3: that the challenges of mixed home/office working, particularly for new staff, continue to be considered and addressed.

In terms of psychometric expertise, the College is well served by their current psychometrician, and additional input from an independent psychometrician, Dr Gay Fagan. I do not believe that further technical expertise of this kind is required in-house although outside experts may always be brought in for particular purposes.

5.3 Style Guide for SBA MCQs

The College has a style guide for the production of SBAs, and I was provided with a copy. However, I heard comments that the style guide is not universally followed perhaps due to it not corresponding to current practice. A unified style guide should be adopted for all multiple choice assessments, including those used for formative purposes (see also section 6.4.1.1). The National Board of Medical Examiners guide to MCQ writing[8] is widely used as a basis for such style guides, with suitable adaptation to particular environments. Once the style guide has been finalised, it should be closely adhered to. Variations in style can contribute to items performing poorly and perhaps having to be removed from the exam, which must be avoided wherever possible.

Recommendation 4: The style guide should be revised to ensure it aligns with common practice, and then it should be used consistently across the College for the development of both formative and summative items.

5.4 Selection and recruitment of examiners

A number of comments were made to me about difficulty in recruiting the required number of examiners for OSCEs, with MRCEM exams being more of a problem than FRCEM ones. Indeed, instances where there was a 'scramble' to find examiners for OSCEs were described to me. The pool of writers for SBAs was also reported as having diminished to the detriment of the exam bank. I will return to the problem of SBA writing in Section 6.4.1.1.

A survey of examiners and potential examiners was carried out in February 2022. This suggested that altruistic motives such as the training of colleagues, the well-being of trainees, the development of the discipline, and patient benefit were significant promoters of recruitment, with financial rewards not being high on the list of potential drivers.

The general problem of insufficient number of examiners is a widespread one for medical Royal Colleges. When clinical staff are placed under extreme stress in their workplace, they may have less time for additional activities such as being an examiner. Perhaps consideration should be given to widening the potential pool of examiners. Currently, it was described to me that the requirements for being an examiner were that the individual was GMC registered, had themselves passed the relevant exams, had held a substantive post for two years, and were active in education. However, for OSCEs in particular, an individual who has expertise in a particular skill may be a very effective examiner at a station reflecting that skill, even if they have not met all

the requirements. That an individual has actually passed that exam or set of exams themselves is a requirement of prestige, not competence. There may be circumstances, for instance, in which a SAS doctor in a senior practice role might well be a most effective assessor. It is a common practice in medical schools to think rather of how well qualified someone is for a particular role in being an OSCE examiner at a particular station than what their general background is.

It was also commented that there appeared to be increasing difficulty in persuading NHS employers to release staff for training in general and examining in particular, which meant that many examiners were carrying out their college duties in their own time. No individual College can address this issue, but I will return to it in Section 10.

A key question is, therefore, how can the role of examiner be made more desirable? The financial cost of paying examiners was described to me as ruling out this option. However, perhaps a more successful approach would be to argue that being an examiner is highly valued as prestige, a chance for networking and professional development, and as a way of helping others.

The esteem of colleagues is a natural desire for professionals, and indeed, can be a major motivating factor. Perhaps the College should explore ways in which this esteem can be made more visible. Award of titles such as RCEM Senior and/or Chief Examiner, RCEM Examiner and perhaps RCEM Associate Examiner for less demanding roles such as writing SBAs can be formalised by presentation of certificates which could be displayed by the recipient. I heard that having a formative SBA item published in the training materials is considered positively by contributors as being capable of being referenced, and perhaps contributing summative items could be recognised by certification which could be included in Annual Review of Competence Progression (ARCP) and revalidation portfolios.

There is also a strong personal development agenda associated with being a College Examiner. It creates valuable networks for the examiner with colleagues from outside their normal working environment. Perhaps it could be associated with Continuing Professional Development (CPD) points in a formal scheme to reflect this.

In this regard, by definition, doctors have chosen to work in a caring profession, and altruism is an important driver. Perhaps the option of becoming an Examiner should be clearly badged as an opportunity to help more junior colleagues and, in a wider sense, the NHS and future patients. This clearly emerged in the February 2022 survey as a major factor in promoting engagement with the College.

Encouragingly, the survey also identified important barriers to engagement, including a lack of confidence on the part of potential examiners in their ability to operate effectively in this role. Training and education around what being an examiner actually involves may well be helpful in this regard.

Recommendation 5: that the requirements for becoming an examiner be reviewed with a view to extending the pool, providing sufficient information and training about the role to alleviate any causes for hesitation in coming forward, and exploring the factors that may make becoming an examiner more desirable.

6. Assessment delivery methods

6.1. The software platforms

As described above, the information systems and information flow in the College is currently complex and vulnerable. Platforms such as OSANA, TRELLO, LearnDash and Kaizen for portfolios are also used. Of the two assessment platforms being used, concerns were expressed about both. One was described as a small organisation, creating some resilience concerns. The other was described as occasionally developing errors, with the Helpdesk being slow to respond. The assessment platform is where the item summative bank is held. Software issues are discussed above in Section 4.1. It is desirable to hold all item-level data in a single source, and this applies to the formative items (which as I also recommend elsewhere should be developed in parallel with the summative items).

The platform which holds most data on candidates is iMIS, and this was represented to me as being the best candidate for the "single source of truth".

Recommendation 6: that both formative and summative MCQ items are maintained on a single platform (currently the assessment platform), and that once the current changes have been embedded, a search for a single appropriate platform for all assessments resumes.

6.2. Online test centre delivery of assessments

The online delivery of the SBA papers is intended to continue. These are currently delivered via Pearson Vue test centres, with exceptions for candidates unable to travel to the test centre either through disability or lack of availability. This process offers advantages to candidates by generally reducing the amount of travel and accommodation expenses. Most respondents favoured this strategy although there were some comments about variability of provision at different test centres. Issues relating to exam security are considered in the following section.

Face-to-face delivery of OSCEs was preferred by both candidates and examiners wherever possible.

6.3. Unfair means

The College had previously experienced the well-recognised problems with item security in high stakes assessments, with candidates exchanging information and passing it onto commercial sites, and, as a result, the roles and responsibilities of examiners have been more tightly defined, and steps were taken to improve exam security. With the move to online assessment, especially where candidates are working from home, a new set of problems may arise, including real time collaboration between candidates and accessing non-permitted sources of information. This is usually addressed by various methods, such as locking workstations to other programmes, video-reviewing candidates during the exam, and continued checking of candidate behaviour while online. Candidate cheating on high-stakes Royal College exams is unusual, but not completely unknown. Generally, in my experience, the ethical standards of candidates for such exams are high. In any case, cheating has costs for candidates. It imposes a cognitive load which in itself may give greater disadvantage to candidates than any advantages brought by cheating. However, concerns may be expressed particularly by those candidates who do not intend to cheat and fear others obtaining an unfair advantage. Because of this, it is worth considering the use of 'similarity-checking' programmes such as the Harpp-Hogan Index or Acinonyx[9]. It is also helpful to employ checks for the presence of item drift, where the scores for particular items increase over time, suggesting they have been compromised.

Recommendation 7: that the College explores the use of 'similarity detection' software for SBA MCQs and monitors item drift across different exam diets for both SBA MCQs and OSCEs.

6.4 Assessment formats

6.4.1 Written assessments

6.4.1.1 Single Best Answer MCQs

While these are the current favoured method, there is evident difficulty in identifying suitable item writers and in generating the number of high quality items required for an adequate item bank. Lack of sufficient quality items may lead to the removal of an undue number of items on paper review, and this has already been identified as a major factor contributing to the scale of the March 2022 FRCEM SBA problem.

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. This should include the development of formative items used in the training process. Formative items writers should implement the same style guide used for summative items, with advantages to candidates in seeing a familiar style of items. There are also benefits in training 'formative' writers so that they can also write summative items. One item bank should be used, and there is nothing wrong with well written items moving between formative and summative categories, as long as the bank is big enough.

A major difficulty in writing MCQs is in developing four distractors. Clinical situations arise where the requirement for this number of distractors is difficult to achieve. However, the occasional use of only three or even two distractors is perfectly allowable and may even increase item quality. Use of one distractor is not recommended, however.

Item cloning, where non-essential features of an existing high-performing item, including the distractors, are altered[10], is also a helpful strategy for increasing the number of items available. In the case of a Royal College where the clinical discipline is relatively clear, Automatic Item Generation should also be considered[11].

There may be opportunities to share items with other Royal Colleges. There may well be topics which are relevant across several colleges – experimental design and statistics may be examples – and medical schools have greatly benefitted from the existence of the Medical Schools Council Assessment Alliance, which facilitates sharing items and best practice.

Every effort should be made to expand the existing pool of item writers, and examiners in general; and this is discussed in Section 5.4.

Recommendation 8: that item cloning, Automatic Item Generation and, where helpful, reducing the number of distractors for SBA MCQs be explored.

6.4.1.1 Very Short Answers (VSAs)

There is current interest in the use of Very Short Answer (VSA) items[12], where a clinical stem is answered as free text by the candidate, often in the form of one or two words. These 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 in fact include multiple cues). However, while VSAs are of interest, they are more 'difficult' in that mean scores will be much lower than corresponding MCQs, with potentially serious consequences for differential performance. Standard setting must be extensively re-thought under these circumstances. There is also much less evidence on the predictive validity of VSA, and they may pose greater challenges to candidates who do not have English as a first language. VSAs also lack the long history of predictive validity of MCQs in healthcare settings.

I do not recommend their use in the RCEM assessments at the moment, but perhaps the College psychometrician should keep a watchful eye on their progress, for instance, by attendance at professional conferences (see Recommendation 17)

6.4.1.2 Written assessments: Conclusions and Recommendations

The current SBA exams of RCEM are of appropriate length and reliability and in the format currently best in accordance with international best practice.

A question I raised with serval interviewees concerned the purpose of the Intermediate MRCEM SBA. This was described to me as similar in standard to the Membership exam. It was indicated that it was used as a marker for transition between ST3 and ST4 roles in clinical practice, but if it is, indeed, at the same level as the Membership exams, perhaps this should be reconsidered given the general desirability of reducing the assessment burden on trainees, in time, cost and stress. It may also allow a more reflective period of development to occur before moving to consultant level posts.

It may be that there is a clear purpose to the Intermediate exam and clear definitions of its appropriate standard in distinction to the Membership and Fellowship exams, but if that is the case, it was not clearly articulated to me. Given the burdens that exams at this level place on both candidates and the College, I make a recommendation that the purpose of the Intermediate exam be reviewed.

Recommendation 9: that the function and standard of the Intermediate MRCEM SBA Exam be reviewed in conjunction with examiners, trainers and trainees.

6.4.2. Practical assessments

6.4.2.1 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 RCEM assessments. The MRCEM and FRCEM OSCEs have been

through a variety of changes, and recent changes in the number of stations were well supported by psychometric evidence and were approved by the GMC. One message emerging from candidates was that a period of stability would be welcome.

There is, however, one aspect of the OSCEs that would benefit from review, and that is the role of the resuscitation stations, and I return to this in section 6.5.2.

6.4.2.2 Practical Assessments: Conclusions and Recommendations

The current pattern of skills measurements via OSCEs is appropriate to the task and of suitable standard setting methodology. However, the collective reliability of the resuscitation stations should be explored by means of a Generalisability study to ensure that they can bear the weight of being a point of failure for the OSCE as a whole, and this is made a Recommendation later in this report, in Section 6.5.3.

6.5 Standard Setting

6.5.1 Methods employed

The Modified Angoff Method is used for standard setting all SBA papers. This is a widely used and accepted form of standard setting. For the MRCEM Primary SBA, the Angoff score is used as the pass mark. For the MRCEM Intermediate SBA and the FRCEM SBA, one Standard Error of Measurement (SEm) is added to the total Angoff score to give the pass mark. The SEm is a measure of reliability of the exam as a whole and adding it to the pass mark reduces the risk of someone passing the exam who does not deserve to pass. Adding an SEm is a common practice, especially in high stakes healthcare licencing exams. However, I have added further comments on this practice in Section 6.5.3.

The MRCEM and FRCEM OSCEs are marked using domain marking. The pass mark for the exam is determined using the Borderline Regression method, as is widespread practice in high stakes healthcare OSCEs. One SEm is added to the pass mark in both exams. However, in order to pass the FRCEM OSCE, candidates also have to pass one of three resuscitation stations.

The number of stations has recently been reduced to 16, plus 2 rest stations, with the concurrence of the GMC. This reduction was supported by a Generalisability Theory Decision study, which confirmed that this did not affect the reliability of the OSCE as a whole.

6.5.2 Removing questions from the exam after it has been delivered

It is a conventional process subsequent to test delivery to both look at the performance of the test as a whole and of each item in a test. This is intended to check that the whole test has performed as desired and, for individual items, to identify any errors (such as the wrong answer having been keyed in at some part of the process) and any items which have performed very poorly.

In Classical Measurement Theory, the reliability of the test as a whole is generally described by Cronbach's α. This is a measure of internal consistency. It ranges from 0 to 1, and we would expect a high stakes exam to have a Cronbach's α of at least 0.7, and ideally, over 0.8. Another (slightly better) measure of reliability is the Standard Error of Measurement (SEm)[13].

The item review may look at how easy or difficult an item is (the Facility of the item, that is, the percentage of candidates who got it correct) and how well it discriminates between candidates (the Discrimination Index and/or Point BiSerial). The Discrimination Index and Point Biserial Correlation are similar, both ranging between -1 and +1, and both compare how candidates performed on that item compared to how they perform on the exam as a whole. If candidates who performed well on the whole test also tended to get that item correct and candidates who performed poorly on that item tended to perform poorly on the whole test, then the item will have a high Discrimination Index and Point Biserial. The individual Discrimination Indices then contribute to an overall discrimination index for the exam as a whole.

An item may cause concern if its Facility is close to 20%. This is because completely random guessing on a 'one best of five' MCQ will give the correct answer 20% of the time. A Facility well below 20% is more likely to be an indicator of incorrect keying of the responses than of most candidates having confidence in a wrong answer. Such an item will also have a low Point Biserial and Discrimination Index (e.g., less than 0.15). However, if the Facility of an item is close to 100%, it will also have low Point Biserial and Discrimination

indices, and this does not necessarily call the item into question. It may be both important and easy, and items should generally not be removed if they fall into this category.

The exam review could also examine how the various distractors had performed, for instance, by means of the Horst Partial Knowledge Index (PKI) which compares the highest performing distractor or distractors with the key. This can be useful in identifying if one of the distractors is very nearly as good as the key, in which case a point can be awarded for both options.

Each time an item with low discrimination is removed from the exam, Cronbach's α, and the discrimination index of the exam as a whole generally increase. In other words, after poorly performing items have been removed, the exam is in a sense a better measure than before. However, this procedure should always be used with caution. In the simplified account I gave in Section 3.3.1, I indicated that the percentage cut score would probably be similar before and after removal. However, it may not be exactly the same. Each time an item is removed, its Angoff probability is also removed. This slightly changes the percentage cut score of the exam. The exam might be better after the removal of poorly performing items, but it is also different.

In general, the average performance of candidates improves slightly when poorly performing items are removed. This is because normally the poorly performing items have low Facility – in other words, many more candidates got the item wrong than got it right. But this is why great caution should be exercised when removing items with high Facility. For these, most candidates got the items correct. In extreme cases, even the rank order of candidates may change.

In the March 2022 FRCEM SBA, 11 items were removed. In my experience this is a high number for an exam of this size. Typically, either no items or only one item is generally required to be removed from high stakes professional exams. The number of items removed in this case contributed to the scale of the reporting error, and many more candidates were affected than would have been the case than if only one item had been removed.

Of course, that so many items were removed, while surprising, does not of itself indicate that they were incorrectly removed. They may indeed have been performing poorly, hence their removal was justified. But it would then be a matter of concern that so many poorly performing items had made it through the process of reviewing items when they had been written and through review of the paper during its development. Item removal should be an exceptional event, and my recommendation is directed to this point.

Recommendation 10: that item quality is carefully reviewed for individual items before they make it into the exam bank, that each exam is reviewed carefully for item quality during the design process, and that great caution is exercised before any items are removed after the exam has taken place.

In making this recommendation I am conscious that it has already been largely acted upon, and that removal of items is now a rare occurrence.

6.5.3 Conclusions and Further Recommendations

The standard setting methods used by RCEM are in accordance with common practice. The addition of a requirement to pass at least one resuscitation station is understandable in light of the importance of this task in EM settings but adds a challenge to the reliability of the OSCE as a whole. It could be argued that the reliability of the OSCE circuit as a whole is now dependent on the reliability of these three stations. While in general the reliability of an OSCE increases with the number of stations, the relationship is complex[14], and it is quite possible for three stations to be reliable, as long as a single ability is under test, as is probably the case here. Nonetheless, it would be valuable to conduct a Generalisability study of the reliability of these three stations separately from the remainder.

Recommendation 11: that the reliability of the three resuscitation stations be calculated separately to confirm that they meet the appropriate standard.

Comments and recommendations are made in Section 9.2.2 on the predictive validity of the exams, and the research described there will shed valuable light on the level of the current standard setting processes.

Currently, a Standard Error of measurement (SEm) is added to the Fellowship exam cut scores, but not to that for the Primary Membership exams, and some discussion has taken place about having consistent practice in this area. But I would advise caution in adding an SEm to the Primary Membership exam cut scores. The

difference may seem slight (the SEm, is typically about 3 or 4%, depending on the length of the exam), but even a small difference like this is likely to have a significant effect on differential attainment (see Section 7.2). The common reason giving for adding an SEm is to enhance patient safety by reducing false positives – i.e., those who pass but, in some sense, do not deserve to pass. But if anything, the RCEM exams err slightly in this direction already (see Section 9.1). The Angoff cut score is the best estimate the judges could make of the required level to pass the exam. Adding an SEm could actually increase the number of false negatives – those who fail but deserve to pass. And in a world where there is a shortage of all doctors, including EM doctors, false negatives are also a hazard to patient safety. In view of the necessary involvement of the regulator and the complex implications of such a change, I do not make a formal recommendation on this point, but if it is necessary to have uniform practice on this point, I would rather see the additional SEm dropped from the Fellowship exams than added to the Membership exams. In the end, only predictive validity studies in the workplace can provide evidence on how this issue should best be handled, and I recommend such a study in Section 9.2.2.

7. Equality, Diversity and Inclusion

There is a legal requirement under Section 149 of the Equality Act 2010, to ensure that individuals with protected characteristics do not suffer discrimination, victimisation or other prohibited conduct. It is worth exploring these issues further in this Section with regard to gender, educational background and ethnicity particularly. 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. Although these characteristics are recorded and analysed by the College, I am not aware of a dashboard displaying performance with respect to these categories longitudinally with time. This would be helpful in spotting trends and identifying the impact of changes currently being made in the hope of addressing these issues.

Recommendation 12: that an EDI 'dashboard' be constructed, so that senior staff can monitor progress on addressing EDI issues on a readily understood graphical basis.

7.1 Gender

On the basis of GMC data covering the period 2014-2019, female candidates did slightly less well than males on the RCEM written assessments but performed better on the practical assessments. In high stakes medical assessments, it is not uncommon for women to perform better than men in professional assessments[15]. Ongoing data to which I have been given access suggests that the sex difference in knowledge tests may be diminishing, while a differential in performance in practical assessments still remains.

Candidate representatives noted that 'less than full time' training was now a possibility and that this was a significant benefit to candidates with families.

7.2 Ethnicity and educational background

It is helpful to distinguish between *differential performance* and *differential attainment*. The former relates to the score obtained, and the latter to the grade this corresponds to (e.g., pass/fail). A small difference in score can lead to a large difference in differential attainment, exacerbated if the mean scores are close to the cut score. If the mean scores actually straddle the cut score, the effect is even more dramatic.

In common with other Royal Colleges, candidates who trained outside the UK performed less well on both written and practical RCEM assessments than those trained within the UK. RCEM candidates identifying as 'white' performed better in both written and practical exams than those from other ethnic groups. Available GMC data runs from 2014-19, but examining the performance of individual recent RCEM exams, I conclude that these differences persist.

Differences in ability between peoples of different cultural backgrounds are likely to lie in socioeconomic and societal factors, including hidden implicit biases in selection and assessment structures. A recent work

programme[16] 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.

Helpfully, the College analyses performance by 'English as First Language' (EAFL) and 'English Not First Language' (ENFL) candidates. For one recent RCEM exam, for instance, there is a differential performance gap of only +1.3% between EAFL and EASL candidates but a differential attainment gap of +4.8% for EAFL candidates.

This suggests that English language proficiency is a significant factor in differential performance and differential attainment. These analyses should be part of the dashboard indicated above.

MCQs show lower 'differential attainment' problems than Constructed Response Questions, especially for candidates for whom English is not the first language[17]. The current use of MCQS, therefore, may help address differential performance gaps between candidates of different cultural backgrounds, and caution should be used before moving to more language based approaches such as VSAs.

The GMC is currently working with Royal Colleges in an attempt to address differential attainment. I have seen a copy of the recent joint RCEM and GMC Medical Royal College and Faculty Action Plan on this issue, which proposes a number of actions to address the challenge. A summary dashboard would be of help in monitoring progress in this regard.

7.3 Reasonable adjustments and Equality Impact Assessments

It would be helpful if there was a single source of guidance on reasonable adjustments for candidates even if each case is considered on an individual basis. Carrying out Equality Impact Assessments (EIA) in the case of exam changes which might affect candidates with reasonable adjustments (and indeed other protected characteristics) is not *required* by the relevant legislation but is nonetheless very helpful in ensuring that due consideration has been given to equality issues. Where significant assessment changes are planned, I recommend that EIAs be routinely carried out.

Recommendation 13: that Equality Impact Assessments be conducted with regard to significant planned changes in the RCEM assessment processes.

8. Training available to candidates

The College has a structured eLearning approach, where candidates who join sign up on iMIS, where an eLearning account is created and linked to the creation of a portfolio, which in turn is operated through Kaizen. Twitter, Facebook, Instagram and LinkedIn are all also used for training purposes. Two podcasts and two blogs are created each month, and there are quizzes and other materials posted for trainees. 'Writing days' are held to generate the required material. Formative self-test items are generated and stored in Trello, a content writing board, which permits discussion between writers and editors. Items are delivered to candidates via a WordPress plug-in known as LearnDash, which offers candidates guidance as to whether their answers were correct or incorrect and collects data on item performance.

Interestingly, there appeared to be a formative 'question writing' strand which was quite separate from that employed for the writing of summative items. Item writers are recruited more widely than for the summative items, with trainees being encouraged to contribute, and international Zoom sessions are held with participants from all over the world. I understand that it is considered prestigious to write items for the College which are then posted on the platform, with this being regarded as something akin to a rapid publication.

Since finding sufficient writers to generate items for the summative bank appears to be a problem, with quite a limited pool of individuals taking part, I found it surprising that the bank and delivery mechanisms were duplicated on different platforms on the learning and assessing sides of the College, and with a quite different pool of individuals employed on each side. It is also of interest that formative item writers regarded the task as prestigious and volunteered from all over the world. I have made a recommendation with regard to the integration of these systems in Section 5.3.

9. Validity of assessments

9.1. Assessment validity

Validity, though a complex and even contested construct, generally relates to the question of whether or not an assessment measures what it is intended to measure[18]. Excellence in standard setting is not in itself evidence of external validity. The real (and often unaddressed) question in healthcare assessment is whether or not the assessment matches and predicts actual performance in the workplace. Of course, there are ongoing workplace-based assessments also aimed at addressing this question, but these are often variably implemented, and, since they frequently have a formative purpose, they may not be implemented with the same rigour as a properly designed summative assessment. Workplace-based assessments are also subject to the phenomenon of 'failure to fail': where an assessor who works closely with a candidate may feel that they ought to fail but is reluctant to do so for a variety of personal and institutional reasons[19,20].

One of my questions to respondents was whether or not they felt the RCEM exams made for better doctors. The answer was generally yes. It was felt that the existence of the exams promoted reading in the field and helped develop background knowledge. In terms of the sensitivity and specificity of the exams, there seemed to be a positive relationship described by trainers between who they expected to pass and who they expected to fail, with a slight bias towards unexpected failures – candidates who they had expected to pass based on their performance in the workplace, but who nonetheless failed the exam. There were rather less frequent accounts of the opposite, where candidates who might have been expected to fail nonetheless passed. This was sometimes ascribed to the candidate having good declarative knowledge but being less comfortable in the very high stress environment of a working EM unit.

In the longer term, there may be discussions at inter-college levels on the place of Royal College exams, and whether they might be replaceable by workplace-based assessments at some point. These discussions can only be resolved by evidence, and in the next two sections, I recommend the gathering of such evidence as the best way of validating the current exam practices of RCEM, and of informing future discussions on the role of exams in higher training.

In the following Sections, I recommend research studies to establish the concurrent and construct validities of the RCEM assessments.

9.2 Validity Research

The meaning of 'validity' in assessment is complex but essentially can be summarised as 'do the exams measure what you want them to measure?' Here I will focus on two relatively simple interpretations of validity as concurrent and predictive validity.

9.2.1. Concurrent Validity

How well do the assessments match the experiences and judgements of trainers and educational supervisors?

As described above, I posed this question to a wide variety of respondents. It may be helpful to view assessments as screening tests. In these terms, therefore, the RCEM exams may be viewed as slightly more sensitive than specific. In other words, those who pass are generally viewed as having deserved to pass, with a few exceptions, but a very slightly 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 also medical and societal costs. My account here is based on qualitative data, and it would be valuable if this could be quantified.

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' noted above. Rather a specific research study should be conducted to compare the confidential estimation by trainers and supervisors of candidates' capability in the workplace with the performance of those candidates in the various exams. This approach would empower the process of standard setting, which at the moment takes place largely independently of evidence on actual performance in the clinical workplace.

Recommendation 14: that a concurrent validity study be conducted to compare performance in the workplace as estimated by trainers and supervisors with performance in the RCEM assessments, with findings incorporated into the ongoing assessment design process.

9.2.2 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 RCEM assessments with ARCP outcomes and Fitness to Practice outcomes through the UK Medical Education Database[21], 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 [22]).

Recommendation 15: that a predictive validity study be conducted to compare performance in the workplace as estimated by trainers and supervisors with performance in the RCEM assessments, with findings incorporated into the ongoing assessment design process.

10. Co-operation with other Royal Colleges and the psychometric community.

Many of the general assessment challenges faced by RCEM are also shared by other Royal Colleges. However, there appears to be no common forum in which these are discussed in the way that such issues can be explored through the Medical Schools Council Assessment Alliance. The kinds of analytic statistics employed, and the values which might trigger investigation or concern, assessment design and delivery issues, problems of exam and item security, detection of possible use of unfair means, and so on are common to all the Colleges. There may even be scope for sharing items. Topics such as statistics and experimental design may have similarities between at least some Colleges.

At an even higher level, questions such as what the assessments are for, and how well they achieved their goals of promoting patient safety and ensuring career progression through the individual specialities require an overview common to all Royal Colleges, if major differences in practice are not to develop. The question whether exams are 'open book' (where the 'book' may include access to Internet resources) is also relevant to an age in which doctors are able to access a wide range of digital resources while they are in actual practice, and it would be fruitful to consider this issue collectively.

Similarly, all Colleges are facing problems with current stresses of delivering medical care in an overstretched healthcare system and of dialogue with NHS bodies about the value of training in general, the release of both trainers and trainees for training and assessment and developing better understanding of examiner roles would be fruitful on an intercollegiate basis.

Differential attainment and performance remain a concern, as described above, and sharing information about successful initiatives with respect to assessment would also be valuable.

Obviously, I cannot make a recommendation to the Academy of Royal Colleges as a whole as part of this review. But I can recommend that RCEM play a full role in initiating and supporting any such joint discussions, where a cross College approach might be useful.

Recommendation 16: that RCEM plays a role in initiating and supporting any cross College initiatives to share assessment expertise and methodologies, including ways of exploring common problems.

There is a wider community of interest in assessment of course than just the Royal Colleges, and this is reflected in a variety of conferences and symposia. It would be of great value if key members of the assessment team were able to attend these events to stay abreast of new developments in assessment in a time-economic manner. The Association for the Study of Medical Education and the Association for Medical Education in Europe run annual conferences at which assessment always features strongly, and the Ottawa

Biennial Conference is devoted to assessment and evaluation. In addition, the internal research carried out within the College on assessment issues such as differential attainment is also of generalisable interest and would, in turn, benefit the wider community.

Recommendation 17: that the College psychometrician and Head of Exams, and perhaps others, are enabled to attend key professional events in the field of assessment and are encouraged to publish research relating to their work.

Summary of Recommendations

Recommendation 1: that a structured appeal process be introduced, so that all appeals and their outcomes could be analysed with a view to identifying procedures in need of improvement.

Recommendation 2: that an integrated system of error management be introduced, encompassing both internally reported errors and externally reported challenges, such as complaints, appeals and surveys, and that communication methodologies such as SBAR be promoted positively to staff and officers to enable the processes of reporting mistakes and challenges.

Recommendation 3: that the challenges of mixed home/office working, particularly for new staff, continue to be considered and addressed.

Recommendation 4: The style guide should be revised to ensure it aligns with common practice, and then it should be used consistently across the College for the development of both formative and summative items.

Recommendation 5: that the requirements for becoming an examiner be reviewed with a view to extending the pool, providing sufficient information and training about the role to alleviate any causes for hesitation in coming forward, and exploring the factors that may make becoming an examiner more desirable.

Recommendation 6: that both formative and summative MCQ items are maintained on a single platform (currently the assessment platform), and that once the current changes have been embedded, a search for a single appropriate platform for all assessments resumes.

Recommendation 7: that the College explores the use of 'similarity detection' software for SBA MCQs and monitors item drift across different exam diets for both SBA MCQs and OSCEs.

Recommendation 8: that item cloning, Automatic Item Generation and, where helpful, reducing the number of distractors for SBA MCQs be explored.

Recommendation 9: that the function and standard of the Intermediate MRCEM SBA Exam be reviewed in conjunction with examiners, trainers and trainees.

Recommendation 10: that item quality is carefully reviewed for individual items before they make it into the exam bank, that each exam is reviewed carefully for item quality during the design process, and that great caution is exercised before any items are removed after the exam has taken place.

Recommendation 11: that the reliability of the three resuscitation stations be calculated separately to confirm that they meet the appropriate standard.

Recommendation 12: that an EDI 'dashboard' be constructed, so that senior staff can monitor progress on addressing EDI issues on a readily understood graphical basis.

Recommendation 13: that Equality Impact Assessments be conducted with regard to significant planned changes in the RCEM assessment processes.

Recommendation 14: that a concurrent validity study be conducted to compare performance in the workplace as estimated by trainers and supervisors with performance in the RCEM assessments, with findings incorporated into the ongoing assessment design process.

Recommendation 15: that a predictive validity study be conducted to compare performance in the workplace as estimated by trainers and supervisors with performance in the RCEM assessments, with findings incorporated into the ongoing assessment design process.

Recommendation 16: that RCEM plays a role in initiating and supporting any cross College initiatives to share assessment expertise and methodologies, including ways of exploring common problems.

Recommendation 17: that the College psychometrician and Head of Exams, and perhaps others, are enabled to attend key professional events in the field of assessment and are encouraged to publish research relating to their work.

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. He served on the Expert Reference group for the GMC's Applied Knowledge Test, part of the proposed national Medical Licensing Assessment, and is currently 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 healthcare settings. In 2022, he was awarded the Gold Medal of the Association for the Study of Medical Education for his services to the field.
The author has no personal or financial conflicts of interest with regard to this review.

References

[1] Senders, J.W. and Moray, N.P., 2020. *Human Error:: Cause, Prediction, and Reduction*. CRC Press.
[2] Rodziewicz, T.L., Houseman, B. and Hipskind, J.E., 2022. Medical error reduction and prevention. *StatPearls [Internet]*.
[3] Waterson, P., Jenkins, D.P., Salmon, P.M., Underwood, P. 2017. "'Remixing Rasmussen': The evolution of AcciMaps within systemic accident analysis." Applied Ergonomics 59 (Pt B):483-503. doi:10.1016/j.apergo.2016.09.004.

[4] Larouzée, J. and Guarnieri, F., 2015. From theory to practice: itinerary of Reasons' Swiss Cheese Model. *Safety and reliability of complex engineered systems: ESREL*, pp.817-24.
[5] Klamar, A., Horvath, D., Keith, N. and Frese, M., 2022. Inducing Error Management Culture–Evidence From Experimental Team Studies. *Frontiers in Psychology*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.716915/full

[6] Helmreich, R.L., Merritt, A.C. and Wilhelm, J.A., 2017. The evolution of crew resource management training in commercial aviation. In *Human Error in Aviation* (pp. 275-288). Routledge.
[7] Stewart, K.R., 2016. SBAR, communication, and patient safety: An integrated literature review. https://scholar.google.pl/scholar?hl=en&as_sdt=0%2C5&as_vis=1&q=SBAR+healthcare&btnG=
[8] https://www.nbme.org/sites/default/files/2020-11/NBME_Item%20Writing%20Guide_2020.pdf
[9] McManus, I. C., Tom Lissauer, and S. E. Williams. "Detecting cheating in written medical examinations by statistical analysis of similarity of answers: pilot study." *Bmj* 330, no. 7499 (2005): 1064-1066.
[10] Luecht, R. and Burke, M., 2020. Reconceptualizing items: From clones and automatic item generation to task model families.

[11] Royal, K.D., Hedgpeth, M.W., Jeon, T. and Colford, C.M., 2018. Automated item generation: The future of medical education assessment?. *INNOVATIONS*.
[12] 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.

[13] Tighe, J., McManus, I.C., Dewhurst, N.G., Chis, L. and Mucklow, J., 2010. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP (UK) examinations. *BMC medical education*, *10*(1), pp.1-9.
[14] Brannick, M.T., Erol-Korkmaz, H.T. and Prewett, M., 2011. A systematic review of the reliability of objective structured clinical examination scores. *Medical education*, *45*(12), pp.1181-1189.

[15] 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.

[16] 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.

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

[18] Downing, S.M., 2003. Validity: on the meaningful interpretation of assessment data. *Medical education*, *37*(9), pp.830-837.
[19] Yepes-Rios, M., Dudek, N., Duboyce, R., Curtis, J., Allard, R.J. and Varpio, L., 2016. The failure to fail underperforming trainees in health professions education: A BEME systematic review: BEME Guide No. 42. *Medical teacher*, *38*(11), pp.1092-1099.
[20] 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.
[21] https://www.ukmed.ac.uk/
[22] 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