INDEPENDENT REVIEW OF THE FEDERATION OF ROYAL COLLEGES OF PHYSICIANS 2023 ASSESSMENT ERROR

Professor John C. McLachlan

October 2025

Contents

Executive Summary

On February 19th, 2025, a statement was released by the Federation of Royal Colleges of Physicians of the UK, that incorrect Part 2 Examination results from the September 2023 Written examination had been released to candidates. The error had been identified retrospectively, as part of a separate review of the performance of anchor items in early 2025. Audits indicated that no other examination had been affected. Erroneous outcomes had been given to 283 of 1451 candidates. Of those, 222 candidates undertaking the exam in the UK (202 UK graduates and 20 international candidates) had been told they had passed when they had in fact failed. Sixty-one (all sitting the exam outside the UK) had been told they had failed when in fact they had passed.

The Federation statement contained an apology and notification that affected candidates would be contacted individually. It indicated that after discussion with the General Medical Council (GMC), those candidates who had retrospectively failed would have to resit and pass the exam to progress in their training. Support mechanisms that would be put in place were briefly indicated.

Summary explanations of how the error had occurred were subsequently posted by the Federation, along with confirmation that an Independent Review would be commissioned. This Review is the result of that decision. Its scope included review of the management and delivery of Federation assessment processes, the causes of the error and necessary improvements, and the impact on the affected candidates. The methodology involved gathering evidence from affected candidates by interview and written statements, interviews with Federation and College staff, software providers and other interested parties including the GMC and data analysis of exam outcomes. No document or data request was declined.

From these processes, a clear picture has emerged of how the error occurred, why it was not identified before results were released and went undetected for so long.

In brief, for the third Part 2 Written exam in 2023, a file was sent to Surpass Assessment, the company which remotely delivers the Federation exams, which contained a 'dummy' answer (Option A for all questions). For security reasons the correct options were not entered, but Surpass systems required that an 'answer' be present. After the exam had been administered, the files for UK Remote Online Proctored candidates were returned to the Federation. At this point, the dummy options should have been manually removed, and the correct answers installed. A human error meant this did not happen. Instead, the correct options were uploaded in addition to the dummy answers. This meant some 80% of the questions appeared to have 2 correct options – the true answer and A. All UK candidates, therefore, received a higher 'score' than they should have done. When the error was corrected, their scores decreased and for 222 of them, their score now fell below the pass mark.

Candidates sitting internationally were not initially affected since their scores arrived separately and were not subject to the same error. But since the pass mark is determined by 'test equating', the apparently higher performance by UK candidates erroneously raised the pass mark slightly, enough for 61 international candidates to be told they had failed, when they had in fact passed.

A higher-than-usual pass rate for UK candidates was queried, but when investigations failed to reveal the error, a second error occurred in that the incorrect results were released to candidates. The errors were only discovered due to a further, unrelated, error identified in February 2025, which triggered an audit of past results.

Some 50 affected candidates provided input, either via Teams or by written statements. They gave moving and detailed accounts of how they felt that they had been affected both emotionally and practically by the error. Those who shared their feelings with me reported experiencing distress, some considerable, and others felt that their ability to successfully resit the affected assessment had been compromised. Training was disrupted, particularly by the decision of the Statutory Education Boards to withdraw candidates not yet in Higher Specialty Training from the April application round. Many candidates reported having made life and career decisions which they felt materially affected their ability to respond to the error by resitting the exam, and adverse financial consequences were also mentioned. Some candidates indicated challenging interactions with life circumstances, such as personal illness and bereavement and others expressed feelings of loss of trust, anger and frustration with the Federation and the need to know how the error had occurred and why it had been undetected for so long.

I have received credible assurances that there have been no further calculation errors in post-COVID Part 1 and 2 Written exams, and where I have been able to check data independently, my findings are consistent with this assurance.

Using Reason's Root Cause Analysis framework[1], the first error was a skill-based error, and the second was a decision-based error. However, in my view the root cause of both these errors was an organisational error in culture, process and resource management at Federation level. The voting members on the Board are the 3 College Presidents and a 4th College representative, while the Chief Executive Officer, the Chief Operating Officer and the Executive Medical Director of the Federation are ex-officio members, as are the three College Chief Executive Officers, but do not have voting rights. The Federation generates an operating financial surplus in which the Colleges share to support their charitable aims. This situation presents a complex task of balancing interests, which, it was indicated to me, led to delays in implementing necessary changes and updates to the exam system infrastructure.

In my view, secondary causes of the first error were the complexity, opacity and vulnerability of the exam processes. Secondary causes of the second error were a lack of quality assurance mechanisms, complacency that 'everything had been fine in the past', and some challenges in working relationships within the operations of the Part 2 Board.

For both errors, there were system stressors, in the form of major changes in exam delivery following COVID, increasing numbers of candidates, pressure for rapid release of results, and the paucity of suitably qualified staff. Previous abnormal results in 2017 caused by an 'escape' of items to some candidates led to this being suspected as the cause of the high pass rate in 2023. However, no evidence was found of item escape in 2023. By chance, the 4th exam in 2023 had an apparently genuinely higher than usual pass rate. These factors led to reassurances being given to Exam Board members that the 23/3 pass rate was part of normal variation.

Both errors remained undetected until February 2025, when a retrospective investigation of exam outcomes prompted by an unrelated issue identified them. Following a series of emergency meetings and consultations with the GMC, the error was made public through the Federation website and social media. In addition to an apology, various remediation steps within the power of the Federation were immediately made available to candidates, including refunds, free resits, removal of the affected attempt from their record, and one-to-one meetings with senior doctors. The GMC accorded with the view that those now known to have failed should be required to re-take the exam. The Four Nation Statutory Education Boards made the decision to withdraw applicants for Higher Speciality Training from the interview process.

I make various technical and organisational recommendations to the Federation. It was indicated to me by senior officers of the Federation, the Colleges and the GMC that there was no realistic prospect of reconsidering the decision to require candidates to resit the Part 2 exam. Instead, I make suggestions as to how such an exam error might be addressed in the future, now that the full impact of the exam error on candidates is clear. I will suggest that decisions should be made on a suitable timescale to allow consideration of all the impacts on affected candidates.

Although one individual made the initial error, this was part of a stressed and flawed system with a number of single points of failure, which made human error likely, if not inevitable. There was no parallel processing in the system. The second error, in releasing the erroneous results with the increased pass rates, was made after checks indicated that no items had been released prematurely. This was the wrong question, but in diagnostic terms, it led to premature closure resulting from availability bias and insufficient consideration of differential diagnoses.

Understandably, affected candidates have called for individual retribution and resignations. However, the current Chief Executive Officer and Executive Medical Director of the Federation were not in place, or only very newly in place, at the time of the errors, and ironically had been working to address the problems in the exam system. The roots of the failure are organisational and structural and go back well beyond the date of the error.

Section 1. Background

While the details in this section may seem dry, they have an immediate bearing on how the error occurred, and I urge readers not to omit it if they wish to gain a full understanding of how the error occurred.

1.1 The Royal Colleges

The Royal College of Physicians, the Royal College of Physicians of Edinburgh and the Royal College of Physicians and Surgeons of Glasgow all have long histories going back to the 16th century. The Colleges fulfil a variety of roles including giving advice and making representations to Government on health issues. The three Colleges independently deliver the Part 2 Practical Exams (PACES) in the UK, although the exam is developed by the Federation.

1.2 The Federation of Royal Colleges of Physicians

The Colleges jointly mandate the Federation of the Royal Colleges of Physicians, which is not a legal entity, but is established by a Memorandum of Agreement between the Colleges.

The Federation develops and delivers the Part 1 and Part 2 Written Membership and PACES Examinations. It also delivers eleven Specialty Certificate Examinations (SCE) and further subspecialty exams. The Federation delivers PACES internationally. The Federation also delivers Continuing Professional Development and Training.

1.3 Federation Management Structure and Governance

The Board of the Federation makes strategic decisions concerning the Federation operations. The three Presidents of the Royal Colleges and one other, currently the Treasurer of the London College, have voting rights. The Chief Executive Officer, the Chief Operating Officer and the Executive Medical Director of the Federation along with the CEOs of the colleges and the Federation Medical Directors for Training, Assessment, CPD, International Training and Development and International PACES are ex-officio members, but do not have voting rights. Operational matters are dealt with by the Federation Executive Management Committee (FEMC), composed of the CEOs of the Federation and Colleges, and the Executive Medical Director and Chief Operating Officer of the Federation. FEMC reports to the Board.

Each of the four major activities within the Federation has a Management and Policy (MAP) Board and a Director. For training, this is the Joint Royal Colleges of Physicians Training Board, JRCPTB, for continuing professional development the CPD Board, for International Development and PACES, the International Board, and for Examinations, the Membership of the Royal Colleges of Physicians of the UK, MRCP UK Board. These report to the Federation Senior Leadership Team (SLT). The MRCPUK Examinations strand will be the focus of this review. Both Part 1 and Part 2 Written Exams and PACES have an Exam Board with a Chair and Secretary, which reports to the relevant MAP Board.

Up till 2023, the Federation had a Chief Operating Officer (COO), but not a Chief Executive Officer (CEO). The COO was at that time managed by the London College CEO. After September 2023, the Federation had both a CEO and a COO.

1.4 The Federation Examinations

The Membership of the Royal Colleges of Physicians of the United Kingdom (MRCP UK) administered by the Federation is a postgraduate medical qualification essential for physicians training in medical specialties in the UK. The Federation exams are part of the postgraduate medical training programme approved by the GMC for this purpose. They are also widely accepted internationally. In total, some 26,000 candidates are examined each year, over many different countries. They are intended to play a role in promoting standards of knowledge nationally and internationally, including for some doctors who cannot access other formal training.

The MRCP UK examination is divided into three parts: Part 1 Written, Part 2 Written, and Part 2 Clinical (PACES).

1.4.1 Part 1 Written

To be eligible for Part 1, candidates must have completed their primary medical qualification and be registered with the GMC or an equivalent body. International candidates must ensure their medical degree is recognized by their home country's regulatory authority. Candidates must have a minimum of 12 months' postgraduate experience in medical employment. The exam assesses knowledge of clinical sciences relevant to medical practice, including cell biology, clinical anatomy, physiology, pharmacology and pathology.

1.4.2 Part 2 Written

The Part 2 Written exam can be taken by physicians in training who have passed the MRCP(UK) Part 1 exam. It builds on the knowledge assessed in Part 1 and tests medical knowledge, skills and behaviour as specified in the UK Speciality Training Curriculum[2] and the Curriculum for Internal Medicine[3]. It evaluates the ability to apply clinical understanding, interpret clinical information and solve problems, including prioritising diagnostic or problem lists, planning investigations, selecting immediate and long-term management plans and assessing prognoses.

1.4.3 Part 2 Clinical PACES

PACES (Practical Assessment of Clinical Examination Skills) is the final stage and tests the ability of candidates to carry out clinical examination skills in a structured setting[4]. PACES currently consists of a half-day examination and includes five stations where there are either patients with a given condition or trained simulated patients. Seven core skills are assessed, and there are a total of 8 patient encounters assessing 7 skills over the 5 stations.

1.4.4 Speciality Certificate Exams[5]

Specialty Certificate Examinations (SCEs) are a postgraduate qualification administered by the Federation. Their aim is to offer physicians a way of demonstrating to prospective employers they have the necessary knowledge and skills to hold a specialty post and are a compulsory part of a Certificate for Completion of Training (CCT) for all UK-based resident doctors pursuing a career in eleven specialties: Acute Medicine, Dermatology, Endocrinology and Diabetes, the European Specialty Examination in Gastroenterology and Hepatology (ESEGH), Geriatric Medicine, Medical Oncology, the European Specialty Examination in Nephrology (ESENeph), Neurology, Palliative Medicine, Respiratory Medicine and Rheumatology.

1.4.5 Delivery

The written exams are currently delivered via Remote Online Proctoring (ROP) for candidates sitting in the UK, and at international test centres as Computer Based Tests, but increasingly also by Remote Online Proctoring internationally. Each College delivers their own PACES, while the Federation delivers international PACES, and this leads to some differences in the administrative processing of candidate information and performance. Marking is via Optical Mark Sheets, which are currently then scanned by Federation staff, but use of tablets is intended to replace this system in time.

Both Part 1 and Part 2 Written exams are delivered as computer-based tests with multiple-choice questions in a 'one best of five' format. Candidates answer 200 items across two papers, each lasting three hours. It is intended that questions are delivered to candidates in random order to reduce the possibility of collusion during exams.

1.5 Data Flow for Part 1 and 2 Written Exams

The question bank is stored in commercial software (risr Assess[6]) then the exam is manually extracted to separate documents by exam editors who create two draft 100-item papers, plus a spare paper, for each exam, in the form of Excel spreadsheets. These then pass to the relevant Exam Board, and Board members are sent items, with one expert and one non-expert reviewing each item. The Board meeting then reviews the papers in their entirety at a 2-day Teams or hybrid meeting, at which item currency and facility are considered. Items identified as potential anchor items (see Section 1.6) are also reviewed at this meeting. A business meeting is part of this emendation meeting, at which previous exam results are discussed. Once the paper content has been agreed, the papers are placed in a secure SharePoint along with candidate lists in the form of an Excel spreadsheet, and there they are accessed by staff from Surpass Assessment[7]. These files do not contain the correct responses to the items, for security reasons. Instead, a 'dummy' answer (the option 'A') has been inserted, since the Surpass software requires an answer to be present. Surpass delivers the exams remotely, within the UK by Remote Online Proctoring, and to the international test centres. The results are returned to the SharePoint, as an Excel or CSV file, and from there, they are extracted by Federation staff from the slightly anomalously named 'Research Unit' and transferred to SPSS files. At this point the dummy answer should be removed and the correct options added. Test equating and calculating the pass mark is calculated in the Winsteps[8] programme (See Section 1.6). Candidate outcomes are converted to a scaled score (rather than a percentage correct score) through this process and are transferred via Excel files to a Java database, prior to release to candidates.

1.6 Standard Setting

Standard Setting is the process of determining the pass mark for high-stakes exams. These exams inevitably vary slightly in difficulty at each administration, and standard setting is intended to ensure that a consistent standard is nonetheless present across all sittings. A variety of standard setting methods are available, such as Angoff, Ebel and Hofstee approaches[9] and previously the Federation used a combination of Angoff and Hofstee methods. The method currently used by the Federation[10] is test equating using anchor items. These are items which have been used before, and have performed well, and at a known standard. A proportion of anchor items (for example, 40%) is present in each exam, and is used to establish the level of difficulty of the remaining items, which may be completely new or drawn from the bank without being anchor items. The difficulty of the anchor items is used to infer the difficulty of the new or non-anchor items and hence deduce a level which is consistent from exam to exam. Only the performance of UK candidates is considered in the standard setting process, but the outcome is used for all candidates. The mathematical process underlying test equating is rather complex, requiring the use of Item Response Theory, and the calculations are carried out in specialist software (Winsteps) which is not readily accessible to non-experts. As a result of the process, a scaled score is calculated for candidates and this is the information returned to candidates, rather than a percentage score.

1.7 Results Release Governance

As indicated in Section 1.5 above, the Exam Boards review the data from previous exams. However, these results have already been circulated to candidates, and this therefore does not represent 'sign-off' of the results in the way that would normally be considered appropriate for a high-stakes exam. Instead, the task of determining that the results can be released to candidates was undertaken by the Board Chair alone or with the Board Secretary. However, the role of the Board Chair, who is always medically qualified, is primarily to confirm confidence in the medical content, particularly if items have been flagged for possible removal. If an item has low discrimination (i.e., the ability to distinguish generally high performers from low performers, measured as the Point BiSerial), this is indicated by the Research Unit to the Board Chair, who makes a determination as to whether that item should remain in the exam, or be removed. If, for instance, one item is removed, the exam will then be out of 199 rather than 200. This is a standard procedure in high-stakes assessment.

Although they may be provided with performance data summaries for the items and the exam as a whole, it is difficult for the Part 1 and 2 Chairs to interrogate the source data, and they are largely reliant on the results presented by the Research Unit. As will be seen, this structural weakness contributed to the erroneous results being released to candidates.

1.8 Relevant Prior Events

There had been previous events affecting various MRCP exams which had relevance to the 2023 Part 2 third sit (for brevity referred to as 23/3) error in various ways.

1.8.1 The 2017 'item escape'

In the analysis of the outcomes of a Part 2 Written exam in 2017, an unusually high performance on the part of one group of candidates was observed. Correlation with their Part 1 results was carried out, and, since as indicated in Section 5.7.1 candidates generally perform consistently, it was found that their Part 2 performance was discordant with their Part 1 performance. Sophisticated mathematical analysis was able to establish the very low probability that this was due to chance and, moreover, that only certain questions were affected. Subsequent investigation revealed that a number of items had escaped from a draft paper into commercial study materials, which had quite innocently been used by these doctors. As we will see in Section 4.3, this is highly relevant to the 23/3 error, in which again high performance of a group of candidates was observed.

1.8.2 The 2021/1 ordering error

During the delivery of Part 1 and Part 2 Written ROP exams, the order of questions each candidate sees is randomised to increase the difficulty of collusion. The items subsequently have to be restored to their original 'canonical' order. For a Part 2 Examination in 2021, an error occurred in the archiving of the items, so that the options were incorrectly matched to the outcomes for some questions. Since this affected the archiving only, it did not affect candidate outcomes but did affect the data available when those questions were later used as anchor questions in January 2025. Members of the Part 2 Board spotted that some answer options were

misaligned with the performance data of each option. This misalignment would prove crucial to the detection of the error in the 23/3 Part 2 exam results since it triggered an investigation of past exam outcomes, which by chance revealed the error in the 23/3 Part 2 exam processing.

1.8.3 The 2022 PACES error

During the paper scanning process following a PACES exam in 2022, the optical mark reading machine, whose age had been recognised as a risk factor, broke down irretrievably. A substitute machine was brought in, but the mark sheets could not automatically be read by the new machine, and human and agency staff had to be drafted to read aloud the results for recording. Errors occurred during this process and a total of 269 candidates were subsequently found to have been given the wrong marks, with 11 being given the wrong pass/fail outcomes.

The company which made the scanner had previously advised the Federation that they were unable to service it due to its age. When it finally broke down in October 2022, they were unable to fix it since they no longer had the relevant software.

There had been previous problems with the scanner and subsequent transcription errors in 2017, resulting in 16 candidates being given the wrong marks, and in one case the wrong pass/fail outcome. This was reported to the GMC in 2018. The Board had at this point approved investment in a new scanner. However, the purchase was evidently delayed, for reasons that remain unclear.

Section 2. The 2023/3 Error

2.1 How the initial error occurred

The third sit of the Part 2 Written exam in 2023 (23/3) was undertaken on September 6th. Two days later, a CSV data file was sent by Surpass Assessment, the company which delivers the remote assessment, to the Federation. Since, for security reasons, the correct answers are not given to Surpass by the Federation, the file contained 'dummy' answers, in which option 'A' was selected as the answer for all questions (the options are labelled A to E).

Normally these dummy answers are deleted before the candidate answers are married to the correct answer key. On this one occasion, due to a human handling error, the dummy answers were not deleted but instead overwritten with the correct answers. If A was indeed the correct answer, no change therefore occurred. If it was not, both option A and the correct option were accepted as correct by the software. Online candidates therefore could receive a mark if they had chosen the correct option, or if they had wrongly chosen option A. This inflated their scores and led to 222 candidates exceeding the calculated pass mark, when if their true score had been recorded, it would have been below this level. This also apparently and erroneously raised the average performance of UK candidates, on which the pass mark was calculated. Consequently, the pass mark was also slightly increased due to the process of test equating.

Candidates sitting the exam internationally were not affected by the error in terms of their raw scores since their data arrived via a different route, and the error was not replicated with their results. But because the pass mark was artificially elevated, 61 international candidates were told they had failed when in fact they had passed.

2.2 Discovery of the 23/3 error and why it took so long

The 23/3 error was effectively discovered only by chance, long after the results had been released. The 2021/1 error described in Section 1.8.1 had led to a misordering of the options for certain anchor questions. When these were examined by the Part 2 Board during their January 30th-31st 2025 meeting, Board members observed discrepancies in their previous performance data. This triggered a re-examination of past Part 1 and Part 2 exams back to 21/1. As described above in Section 1.8.2, while the mis-ordering proved not to have affected candidate outcomes, it was discovered entirely by chance that there was another error affecting the 23/3 exam, as detailed in Section 2.1.

Since it is clear that the discovery of the 23/3 error was accidental, it raises the possibility that (a) it might never have been discovered or (b) it might have been discovered even longer after the release of the results to candidates. This latter consideration will lead me to make some proposals as to how an event with various time lags might be handled in the future.

2.3 Opportunities to identify the error at the time

I have indicated that in my view there were two formally separate errors: first the incorrect calculation of the outcomes and second, the release of the incorrect results to candidates. The system, therefore, failed at two points. After the draft results had been calculated but before they were circulated to candidates, the results contained a number of signals that a problem had arisen, some of which were clear and others more subtle. These are indicated below. The reasons why these signals were missed is explored in Part 4.

2.3.1 The pass rate for UK graduates was markedly raised

The pass rate for UK graduates was 96%, whereas the corresponding average pass rate for the previous 10 exams was 81.2%, with a standard deviation of 4.9%. In other words, the UK pass rate was about 3 times the standard deviation above the previous 10 exams (none of which were more than 2 standard deviations away from the mean). If this had been a clinical measurement, one would say it was well outside the reference range. This elevated pass rate was identified and queried by the Chair of the Part 2 Board, and an investigation took place, but subsequently he was reassured by the Research Unit that no error had been found and that it represented normal variation. As I note in Section 1.7, the role of the Chair is essentially to ensure the quality of the medical content of the exam, and they were not in a position to interrogate the original data itself, in the light of the complexity and opacity of the processes involved in data handling and standard setting. They had little choice but to accept the assurances he received, and the results were released.

2.3.2 The discrimination of the exam (mean Point BiSerial) was affected

How well an individual question discriminates between high and low performing candidates is frequently measured by the Point BiSerial (PBS), or alternatively by the Discrimination Index. From this, the discrimination of the entire exam, the mean PBS, can be calculated. Since there were apparently two correct answers for 80% of the items, one of which was selected effectively randomly, one would expect that the ability of the exam to discriminate between candidates would be affected also, and indeed the mean PBS was markedly reduced for UK graduates in the 23/3 exam. The mean PBS for the 23/3 exam was some 3 standard deviations below the mean value for the previous 10 exams. This is significant because, even if the pass rate was higher than usual, we would not expect the mean PBS to be affected. However, the mean PBS was not routinely calculated for the Federation written exams.

2.3.3 Other affected metrics

Of course, other metrics in 23/3 for UK graduates were also affected by the error. The mean score was almost 5% higher than usual, but the pass rate reflects the score, so this would pass as the same signal. The standard deviation was lower than usual by a marked amount, because the results were more bunched together than usual, and this might have drawn attention had it been routinely reported. The distribution remained normal, though it was shifted higher, rather than spreading over a greater range as might have been expected.

2.4 The time course of events around and subsequent to the discovery

The following dates are drawn from the Federation Internal Review of the error but accord with independent evidence I received from Federation staff and members.

The issue of statistical discrepancies in the anchor questions from 21/1 (see Section 1.8.2) was raised by Part 2 Board members following the Board meeting on 31/01/2025. These were investigated by Research Unit staff from 03/02/2025 until 07/02/2025 since the investigation was broadened to cover all exams since 21/1. This established that the 21/1 error had not affected candidate outcomes. However, the error in the 23/3 results was discovered by chance during this full review and that it had affected the exam pass/fail outcomes of 283 candidates, 222 of whom (202 UK and 20 international) had undertaken the ROP exam and had been told they had passed when in fact they had failed, and 61 international candidates who had been told they had failed when in fact they had passed.

The discovery was escalated to the Federation Chief Executive Office, the Federation Executive Medical Director, the Medical Director for Assessment and Associate Medical Director for Written Examinations on 10/02/2025, and to the College Presidents and CEOs. Work was carried out to establish the identity and stage of training of the affected candidates, and email communication with the GMC commenced on 17/02/2025. In the course of these discussions with the GMC, the Federation accepted that the position was that the pass mark would stand, i.e. that candidates who were narrow fails (within 1 Standard Error of Measurement) would not be considered as a 'condoned' pass. A meeting was convened on 19/02/2025 with the College Presidents

and announcements on the issue were made to affected candidates and other stakeholders on 20/02/2025. The Federation indicated that they were taking the following actions: (a) informing all impacted candidates (b) offering one-to-one support by senior Federation clinicians (c) establishing a helpline (d) refunding examination fees and providing free resits and (e) exploring other help according to individual circumstances.

During these processes and consultations, it became clear that the information had leaked to at least some external individuals, and accounts had begun to spread that something had gone awry.

The Federation advised each of the SEBs about the issues on 19th February. A range of Teams meetings ensued as the number and details of the affected candidates were confirmed. A legal meeting Teams call on 24th February, in the light of information gathered, was held. In these discussions, among the issues explored was the position of potential candidates who had not applied for interviews on the basis that they had not passed MRCP. The final decision, communicated to those affected on 27th February was that affected doctors already in HST posts would be allowed to continue. However, on the basis of legal advice doctors applying for HST would no longer be eligible to progress with their applications, although they would be offered a subsequent interview on passing. The SEBs arranged for Deans teams to contact trainees and ensured LEDs (locally employed doctors) were also covered by working with the Medical Directorate to get guidance and information to Medical Directors about LEDs in their individual Trusts. They ensured people knew they could get additional study leave and educational support if needed and well-being support from Professional support Units across the country.

The SEBs agreed to support trainees in higher specialty training to remain there having been appropriately reviewed by their supervisors to ensure they and patients were safe and that the same was done for LEDs.

It was also agreed that individuals could progress straight to interview when they next applied and looked at where they had applied to ensure posts were held to add in for round 3 so there were relevant posts if they chose to apply

The Presidents of the Colleges wrote to the Statutory Education Boards (SEBs) on 23/02/2025, asking that applicants for Higher Speciality Training (HST) who had interviews should be allowed to proceed. The SEBs replied on 26/02/2025 reiterating the decision of the 20/02/2025 meeting.

A variety of further communications to those affected was published on the Federation website in subsequent days[11]. A key part of the immediate response from the Federation on discovery of the error was focused on communication with relevant stakeholders and particularly the affected candidates. All candidates were contacted with the offer that Federation would arrange a one-to-one discussion for them with a consultant physician with experience of medical training pathways. In support of this, over 20 consultants provided their services and time to listen to the impacted candidates who were able to express their anger, anxiety and upset and receive practical advice about future possibilities. Many of the volunteers found the calls very difficult given the feelings expressed from, and the impacts of the error on, the candidates. The three College Presidents, the Federation Executive Medical Director and the Federation CEO and all involved expressed sincere regret and sorrow that this incident had occurred, and empathy for those affected.

Section 3. Impact on candidates

Over the course of my review, I heard from 50 candidates (some 18% of those affected), either through Teams interviews or via written narratives. These represented all major categories of those affected, including doctors in Higher Specialty and Internal Medicine Training posts, those who had not yet entered training programmes, and international doctors. Although common themes emerged, each case was unique.

A common initial reaction was disbelief. Affected individuals reported thinking it was a mistake, a cruel joke, or a spam email - surely UK Royal Colleges could not make such an error. Only when it began to sink in that it was genuine, did the full range of reactions begin to emerge. Those candidates who were approaching the limit of possible resits reported feeling especially happy to have ‘passed’, and correspondingly, even more dismayed than others when subsequently told they had failed and were concerned about future attempts. Many candidates told me that they had provisionally signed up for the next exam diet, just in case they failed, and would have still been current in their studies, working environment and work/life plans, had they known they had failed. Being told they had passed meant they cancelled these sittings and moved on with their lives and careers.

3.1 Emotional Consequences

Many individuals reported experiencing severe emotional distress as a result of the error coming to light. A number of Interviewees broke down in tears during our interviews, and strong views of shame, humiliation and anger were also expressed.

Doubts of their own capability, brought about by being retrospectively told they had failed an exam they believed they had passed, were common. Others told me that they had felt acute embarrassment at having to tell families and friends, some of whom had attended membership ceremonies, that in fact, one of the exams had been failed.

Several doctors indicated that they were considering leaving the profession and the distress caused by being told of the error meant that many felt less capable of undertaking the Part 2 exam and their other educational and professional commitments than they would have been if they had been told the correct result to begin with. Others felt there was a loss of esteem from co-workers, since they were now publicly badged as 'failures', in ways which they felt may have been less severe if a failure had been recorded at the time of the exam. The publicity surrounding the error meant that the failure was less private than usual and some reported that the media coverage only served to exacerbate their feelings. Others candidly confessed to a feeling of 'imposter syndrome' and were worried that others might think less of them.

The blame for these adverse consequences, respondents felt, lay with the Federation (or 'College', since candidates understandably did not always clearly distinguish between these concepts). A number expressed anger and dismay at the events, and scepticism over future examinations, no longer trusting the Federation to administer future exams competently.

Some affected doctors felt unable to participate in interviews with me, as the mental trauma described as too great, and preferred to send a written account.

3.2 Health Consequences

A number of affected individuals reported serious effects on health, including being signed off work, either temporarily or on an ongoing basis. These were primarily mental health issues, but in some cases involved stress exacerbation of existing conditions. A number of respondents told me that they had required medication and were in some cases still affected. Accounts of distress, of disturbed sleep and distraction from other tasks, were common.

3.3 Life Consequences

Many respondents found the errors were particularly problematic, given their life stage, since it occurred at a time when so many important personal decisions about family, relationships and careers are being made. Affected doctors described how, on the basis of the incorrect results, they had made major life commitments, such as marriages, starting families and house moves, all of which made them feel much less capable of devoting the necessary time to retaking an assessment they legitimately believed they had passed. Those who had started families now have very young children which they felt materially affected their ability to undertake a resit of the exam. Some had planned to start families and were now having to contemplating putting these plans on hold.

Some mentioned that carer status had changed in other ways, with other family members for whom they had responsibility becoming ill, and bereavement involving very close family members was noted, which again made some feel that preparing for an exam would be more difficult, and not like the previous attempt.

Others described stress caused in their relationships, sometimes because their partner was concerned about suffering potential adverse and other career consequences, reflective in my view of the value of partner support during the demanding and stressful time of exam preparation.

The argument has been made, in BMJ[12], that women doctors were disproportionately affected. Several respondents who were single mothers and working less than full time, reported feeling that the added requirement to resit the exam was a particularly challenging burden.

In some cases, those affected told me that they were very worried about their immigration status. Some reported concern that if they failed to pass Part 2, their progression, or even their posts, might be affected. If they were to lose their employed status, they then feared they might be liable to deportation, sometimes to challenging environments from which they had escaped a number of years ago. Several IMGs described a loss of trust, not just in the Federation and Royal Colleges, but in British society as a whole, which they had thought to be fair and trustworthy.

3.4 Career Consequences

The career consequences for candidates depended on the stage of their career and on whether they had undertaken the ROP exams in the UK, or at international test centres. The consequences were of course different for the 222 candidates who were wrongly told they had passed when they failed, and the 61 who had wrongly been told they had failed when they had passed. A view was expressed by candidates in several different categories that the job market was becoming more competitive, and they had suffered in this regard from the long delay in discovering the errors. Some had only undertaken PACES in the belief that they had demonstrated the required knowledge base in the Part 2 exam and felt that they would not have taken this step had they known they had failed Part 2. Others were about to sit PACES and felt that their self-confidence had been damaged by the unexpected discovery they had in fact failed Part 2.

3.4.1 Higher Speciality Resident Doctors

Those who had already progressed into Higher Speciality Training might seem to have suffered less than those still in internal Medicine Training (IMT) who were withdrawn from HST applications. However, in addition to the emotional and life consequences, some had commenced further study programmes such as Masters' Degrees and PhDs, and other specialised courses and professional exams, and were now committed to assessment and submission requirements which some felt would impact on further career progression if not passed or fulfilled.

They reported being placed in the position of having to choose which of their obligations they can attempt. If they focus on retaking the Part 2 exam, concern was expressed that other requirements and assessments might be neglected, with potential harm to their timely future progression, and in some cases, permanent loss of the opportunity. Some felt that failing these additional qualifications or subsequently failing the Part 2 exam again (which was made more likely, they felt, by the stress of the error) might adversely affect their future performance.

3.4.2 Internal Medical Resident Doctors

A particularly acute situation was thought to exist for those affected who had not yet entered HST, but had pending interviews. The Statutory Education Boards decided that such candidates would be withdrawn from the application process until they had passed Part 2, a decision taken at a meeting with legal input, and with the presence of the GMC, although the GMC indicate that they did not provide input to the decision. An offer had been made of a guaranteed interview place after passing Part 2, but this was not felt to be adequate for the disruption to their morale and studies. Like HST doctors, these individuals also felt that the ability of the exam to fairly represent their true abilities was adversely affected by an error which was not their fault.

3.4.3 Other grades

Some doctors had not yet progressed to IMT, remaining, for instance as so-called 'Foundation Year 3' doctors or as other grades. These doctors were closest to the situation of all candidates prior to the exam itself, but still felt that they had suffered all the emotional and life consequences of the error.

3.4.4 International doctors

Sixty-one doctors were wrongly told they had failed, when in fact they had passed. In addition to the emotional and life consequences they reported feeling upon subsequently discovering there had been an error, they advised me that they had often paid for further study programmes, taken irreplaceable annual leave for study and exams, and even made life and career choices based on the belief that they had failed, including the possibility of leaving medicine altogether. Some said that being told they had 'failed' an exam they had cause to feel they had done well in, had caused them to question their own abilities, and in some cases their career choices: it had affected their confidence' or led to demoralisation. I heard of candidates who had accepted other posts than the ones they might have preferred to pursue, had they known they were successful.

While MRCP is not a requirement in other jurisdictions, it has widely been seen as a mark of quality, an advantage in interviews and job applications, and a condition of employment, and some IMGs felt that the credibility of not just MRCP qualifications, but all UK qualifications had been harmed. International candidates had suffered pauses in the commencement of posts, while their circumstances had been investigated further.

Some international doctors reported a feeling of being in situations of jeopardy in their current location and of having hoped to move to the UK as a place of safety.

3.5 Media Coverage

Media coverage fell into two broad categories. Some outlets reported the shock and dismay that candidates felt on learning of the error with headlines such as " 'Catastrophic' error as hundreds of doctors told they passed exams they failed"[13] and "Calls for compensation after hundreds of doctors received wrong exam results in 'atrocious' and 'life-altering' error' "[14]. Others, however, implied that patients had been placed at risk, e.g., "Fury as blunder lets medics stay at work after failing exam" [15] and "Doctors who failed exams working on specialist NHS wards after results blunder"[16]. I was told that these headlines had made some affected candidates feel even worse about themselves.

3.6 Communications from the Federation

General dissatisfaction was expressed about the amount of communication from the Federation, but also sometimes its nature and tone. Those affected often held the view that communications had been inconsistent, for instance, around the number of resits permissible, and when sign up for the March attempt was possible. Others reported that emails had gone unanswered, and that the 'point of care' doctor they had been provided with was remote, and not clear on the current situation. Some drew an analogy with breaking bad news to patients and felt that this could have been handled better.

Section 4 Why the errors occurred

In my view, there were two distinct active errors[17]. The first was the individual human error in failing to remove the dummy options data from the exam process, a *skill-based error* in which a routine task was omitted. The second was more collective: the release of the erroneous data to candidates, despite signals being present and concerns being raised about the signals. This represents a *decision error*, resulting from inadequate information and investigation. Both of these I would consider as the assessment equivalent to medical Never Events[18]. However, I do not consider these the root cause, which is considered in Section 4.1.1.

4.1 The Causative Factors

4.1.1 The Primary Cause

The organisational and financial relationship between the Federation and the Royal Colleges, in my view is the primary root cause of the problem (and was an *organisational level latent error* involving both organisational culture and resource management). In 2023 the Federation was charged with effective and secure delivery of the exams processes but did not have independent control of the resources to deliver the required business systems infrastructure needed, while the Colleges (through the Board) were responsible for strategic decisions, including ultimately expenditure. A complex problem of resource management was therefore created. This led to delays in the timely investment needed to deliver new fit-for-purpose systems infrastructure which would reduce manual processes and thereby deliver more robust and secure exam delivery, with decisions being extensively considered between the Colleges and the Federation before resources were released. The delays to the purchase and deployment of a new scanner in the PACES 2022/3 error, and the delays in the construction of a seamless Exam Management System, are examples cited to me where agreement to expenditure had been made, but this was not the same as timely implementation.

Up till 2023 when the error occurred, the Federation did not even have a Chief Executive Officer, and the Federation Chief Operating Officer was line managed by the CEO of the London College. On the Federation Executive Management Committee in the years up to the 23/3 errors, the Federation COO was therefore outranked by the College CEOs, one of whom was their line manager. It had been obvious for a long time that investment in the exam management systems was required, and that failure to make this investment was hazardous, and indeed featured in the Risk Register: but the actual release of the necessary funds was not made on the urgent basis that was required, and when it was released, there were further challenges around determining an appropriate solution given the complex system requirements which caused further delays. The

appointment of a Federation CEO in late 2023 was a necessary and valuable step, but too late to affect the 23/3 error. However, although this appointment was a step in the right direction, the Federation is still ultimately overseen by the College voting members on the Board. The Presidents are of course elected and have relatively short tenures. They may not have particular expertise in assessment theory or practice and probably make use of input from their respective CEOs. I will return to this issue in my Conclusions and Recommendations.

4.1.2 Secondary Causes

There were a variety of secondary 'latent error' causes for each of the errors, and some 'Swiss cheese' style holes in the processes which aligned by ill chance, and contributed to the inability to detect that the error had occurred before the results were released,

For the first individual human error, which represented a single point of failure, I believe that there were three secondary causes, representing *supervisory factors* (as distinct from *supervisor violations*).

1. The complexity of the Federation assessment processes, which require multiple transitions between platforms and software holdings. Particular problems were experienced at both ends of the exam process, in exam design and in results reporting.
2. Opacity of the Federation processes, whereby even basic information is only extractable by one or two Federation staff members
3. The vulnerability of the processes, whereby only one or two Federation staff members had the knowhow to carry out the scoring process, and only one of them was actively involved in doing so.

For the second error, which also represented a single point of failure, I also believe that there were three secondary causes, also supervisory factors.

4. There was a lack of timely and informed 'Quality Assurance' oversight of exam outcomes before their release to candidates. The Risk Register had identified several challenges to secure delivery of the exam processes, but these had not been converted to urgent action, even though they had been on the Risk Register for extended periods of time. There had previously been an academic quality and quality management committee, but after a reorganisation quality management was the task of a single manager with a large portfolio. The exams MAP Committee reviewed results long after they had had been released to candidates. In the event, release of the results was made on the signature of the Part 2 Board Chair, who, though concerned about the high pass rate, was in no position to verify the accuracy of the results personally and instead had to rely on assurances from the Research Unit. This was not an adequate assurance process.
5. There was a culture of complacency in the Federation in 2023, in that it was believed that 'everything had been going fine' despite the occurrence of repeated errors, some known, others not. The Risk Register identified risks accurately, but this did not lead to them being addressed in a timely manner. Known errors with PACES were considered in isolation, rather than being extended to all exam processes. This was despite common themes (inadequate resources and the vulnerability of manual processes) between the Written exams and PACES. This complacency led to warning signals being missed or indeed dismissed.
6. I heard that working relationships between Federation staff and officers on one hand and some Part 2 Written and SCE Board clinical members on the other, were sometimes subject to tension. This may have led to the adoption of entrenched positions when discussions about discrepancies in exam results were taking place. This is likely to have contributed to the warning signals of the error not being fully explored.

Recommendations to address these causes of error are described in Section 5.

4.2 Stressors in the system

There were a number of factors which added stresses to the exam processes and generally added to the risk of mistakes being made and missed. These are summarised below, and recommendations are made in Section 5. Inevitably, these environmental stressors are interwoven with the specific causes of the errors listed above.

4.2.1 Time pressure

There were defined (and in the circumstances, short) deadlines after each assessment for the release of results to candidates. Meeting these deadlines was seen as imperative. Of course, there is candidate pressure for the release of the results, sometimes exerted through the Colleges and social media, and other deadlines such as the interview round and other exam application deadlines have to be taken into account. However, this strict deadline culture is at odds with careful checking, especially if an anomaly is detected. As a result, results were issued solely on the signature of the Chair of the Board, on the assurance of the Head of the Research Unit, despite the Chair being in no real position to be able to verify the results.

4.2.2 Complexity of Decision-Making Processes

As indicated above the Federation staff were not in independent control of processes and the timely allocation of resources to particular tasks was complex. At times, I was given the impression that this complexity impacted negatively on the ability of the Federation to deliver the best possible assessment processes – for instance, the delay in replacing the outdated scanner, whose failure led to the 2022 PACES error, and delays in implementing a new Exam Management System.

4.2.3 Relationship with Surpass Assessment

The delivery of assessments is subcontracted to Surpass Assessment, and at times there appeared to be a degree of informality in these arrangements which was not conducive to good practice. I was told that some arrangements were verbal rather than fully codified. The decision not to give Surpass the correct answers (on the belief that this added to security) added another step to the process which was either not incorporated in written Standard Operating Procedures or not acted upon.

4.2.4 Complexity

This is identified as a key 'secondary cause' of the initial error. Even by comparison with other Royal Colleges, the processes involved seemed (and to a large extent still seem) complex. Data arrives from Surpass as a CSV or Excel file, is transferred to SPSS where data syntax queries have to be made, standard setting is carried out in Winsteps, and results will return at some point to Excel or CSV files. Each step relies on human manipulation of data. The exam bank is held in a *risr* platform, from which 'editors' extract draft papers to an Excel spreadsheet, and interact with the Exam Boards over reviewing, altering and producing the final papers. These are sent to Surpass without the direct involvement of the team that will assess the outcomes. The correct answers have to be manually input from the question bank to the answer file, and inevitably errors can occur in this process too. I heard that many Federation staff did not fully understand what happened in other parts of the assessment process, and this is hardly surprising. Most of these steps are inaccessible to interrogation by anyone other than a very small number of staff (sometimes just one) with specific and unusual skills. This contrasts with dedicated commercial assessment platforms (a number are available) in which all these steps are generally integrated, and can be readily interrogated and visualised.

4.2.5 Over-reliance on small numbers of trained personnel

While a number of people work on the overall exams process, the number with relevant psychometric expertise is small. The 'Research Unit', which despite its name actually does most of the heavy lifting of results processing, consists of only 3 people, not all at the same level of expertise, and each with unique commitments of their own, meaning that there was little spare capacity for checking each other's work.

4.2.6 Change

Since 2020, and triggered first by the pandemic, there have been major changes in exam delivery methods and software, hardware and practice. Each change has required changes in operating procedures, and these were not always recorded in detail, no doubt due to pressure of work. There was a reliance on tacit knowledge, rather than written records of procedures, and some information (such as some communications with Surpass as indicated above) seems to have only been exchanged verbally.

4.2.7 Expansion

The Federation over recent years has been dealing with Increasing numbers of candidates for the Part 1 and 2 exams, and there was a temporary move to 4 sittings per year. Methods which worked well with smaller numbers of candidates and simpler 'pencil and paper' approaches were replaced with larger scale methods which meant that previous experience was a less good guide to operations.

4.2.8 Change in Personnel and Loss of QA Structures

It was described to me that there had been a number of changes to personnel over the immediately preceding years, with long serving individuals departing for a number of reasons. This included changes to committee structures, and it was indicated to me that some of these may have impacted on quality assurance processes, including the loss of an academic quality and quality management group, and a membership standards and review group. Quality management is now under the remit of the Head of Assessment, Quality & Committee Services which represents a large portfolio for one individual.

4.3 Why were the warning signals missed?

Warning signals in the data (see Section 2.2) did not lead to the data being fully interrogated, or the error being identified either before the release of results, or afterwards at the relevant Board meetings. These signals were not accorded sufficient significance, however, and part of the reason lay in corresponding holes in the 'Swiss cheese' model of medical error occurrence.

The first lay in the 'escape' of items in 2017 (see Section 1.8.1), which had been detected by comprehensive mathematical analysis of the results, particularly the correlation between Part 1 and Part 2 results. In general, candidates in exams perform quite consistently when there are several parts to an exam. The 2017 item escape was detected by some candidates performing in an anomalous manner between Part 1 and Part 2.

The first thought of the statistical staff seems to have been that this had happened again, with items escaping. "The idea had haunted us ever since", one respondent said. However, the same kind of analysis as had been performed in 2017 revealed no sign of this. Candidates had performed consistently between Parts 1 and 2 – in other words, there was a good correlation between the results, with no marked outliers, and the distribution of scores was normal. Therefore, it was concluded that there was no problem with the results.

This is comparable to the cognitive bias in medicine known as premature closure[19], in which the correct diagnosis is not reached because the examination of evidence terminates too soon. This may be influenced by availability bias[20], in which the recency or vividness of a recent case predisposes towards the same diagnosis. The investigation therefore did not proceed further, and the Chair of the Part 2 Board, and subsequently the full Board members, were assured by the Research Unit that the results were not discrepant.

This problem was exacerbated by a second hole in another cheese slice. The results for the 23/3 exam were not presented to the full Part 2 Board until January 2024, three months after they had been released to candidates. At this point, the 23/4 results were also available, and by ill chance, these were genuinely high, with 91% of UK candidates passing. This seemed to provide support for the hypothesis that the 23/3 results were within the normal range of variance. The full Board was still anxious and concerned about the high pass rate in 23/3, and this anxiety continued even after this Board meeting, but there seemed no grounds for acting on these concerns. Of course, if the mean PBS for 23/4 had been examined, it had returned to the normal range, and the 23/3 mean PBS remained discrepant, but this metric was not calculated at that time.

In theory, the next step in the assurance process was that an exam report would go from the Board to the Exams MAP Committee, but again the same reassurances about the high pass rate were made at MAP, and by now this was long after the results had been released.

Section 5. Conclusions and Recommendations

5.1 How confident can candidates be that the amended results are correct?

Affected candidates frequently expressed concern, even scepticism, about the accuracy of what they had now been told, since such a serious error had occurred and gone undetected for so long. I asked for, and received, a variety of original data relating to the error, and previous and past Part 1 and 2 exams. I examined with particular care the full data set for the 23/3 exam, before and after the discovery of the error, and compared it with the corresponding Part 1 results, and other written exams, before and after the discovery. I was given details by candidates of how their scores had changed on notification of the errors, and in some cases, copies of the letters they had received. I could therefore check these against the official files. In all of these analyses, I found no mismatches with the original data set. In analysing past and previous exams, I found no evidence that other errors had occurred with the written exams, beyond those described in this review. It is not logically possible to prove a negative, but I can say that I found no evidence of other calculation errors in the Federation Part 1 and 2 exams.

5.2 How confident can candidates be that future Federation exam results will be correct?

In the aftermath of the error, Federation staff are hypervigilant to the possibility of a recurrence of the same kind of error, in my view making it unlikely in the short term. However, this very hypervigilance might draw attention away from some other areas of vulnerability. A particular risk might involve the introduction of new technology and processes, such as the move to digital results recording of PACES scores, or the introduction of two-camera proctoring. A second might relate to exam security, particularly with the introduction of Remote Proctoring internationally. Concerns were expressed to me about exam production and standard setting in the SCE examinations. And I am still not clear to what extent exam processes from production to results handling are integrated with the development of the new 'Exam Management System' which is being implemented.

In Section 5 I make Recommendations which I believe will help to address some of these risks. Since I had the opportunity to brief senior Federation staff on my interim findings at regular intervals, some actions have already been taken to address these challenges.

5.3 Accountability

Understandably, many candidates expressed a desire that those responsible for the error should be punished. A frequent comment was that as they as individual doctors were liable to be held responsible for their patient errors, so too should those responsible for egregious assessment errors. While it is natural that candidates should have called for such punitive consequences, I nonetheless have not recommended such steps, while recognising that this choice will not be popular with some of those affected. There are several reasons for this choice.

The initial individual human error that was made, resulted from one unsupported individual working under time pressure and financial constraints, with an extremely cumbersome system of data flow with several manual data transfers between different software applications. There was a digital transformation strategy, in view, but this had not yet been implemented, perhaps due to the organisational complexity of the Federation structure and management. The specific detail of the error, that a data set had not been removed, is somewhat analogous to a patient data application set up with dummy patient data, such that every time it was used, the dummy data had to be removed by hand before the genuine data was input. The individual who made the error did this successfully on multiple occasions, but on one occasion, failed to do so. I hope that this analogy may make sense to the affected doctors, who are so often faced with working in time-pressured, under-resourced and unsupported environments. It was the same individual who made the error who subsequently spotted it, admitted their fault and brought it to the attention of the Federation, showing considerable candour.

The second error involved premature closure in diagnostic terms, availability bias, and lack of exploration of differential diagnoses, when the initial diagnosis was disproved. As described in Section 4.1.2, there were a range of secondary causes, with a number of structural and organisational flaws that contributed to the errors, but again stemming from the complexity and opacity of the exam processes.

It is inevitable in such circumstances that there should be calls for the resignation of senior officers of the Federation. But the current CEO and the Executive Medical Director were either not in post, or only just in post when the error was made and ironically have been working to address some of the structural and organisational issues that had caused the errors. The CEO in particular was the subject of much positive comment for their actions prior to the discovery of the error.

Analogies might be drawn with doctors in training, working under pressure and without adequate resources[21]. It is also the case that severity of sanctions can itself pose a safety risk, since it may lead to suppression of information. An analogy is with Never Events such as 'wrong side surgery' in the NHS. These should by definition never happen, but from April 2024 to February 2025 there were 375 Never Events in the NHS. Up till 2018 there had been the possibility of financial sanctions for such events, but this policy was changed, correctly in my view, to one where there were no sanctions. Only if mistakes can be identified and made public, can policies and procedures be changed in such ways as to make the system safer.

If there were one senior single manager or individual who had been grossly negligent, I would identify them. Instead, however, this was a corporate failure, a network of fault, extending over all four organisations, and a number of years. Current Presidents and Chief Executive Officers were not those in post at the time the errors

were occurring, and other individuals involved in the past have retired or moved on elsewhere, even if it was appropriate to assign corporate blame to them as individuals.

Understandably, the affected doctors often expressed negative views about Federation and College officers and staff (not always distinguishing between these bodies, as is quite common). That 'they' were cynical and uncaring about the plight of the affected doctors was a view expressed by a number. I interviewed dozens of Federation and College staff and officers. They used words such as 'catastrophic', 'heartbreaking', 'horrified', 'devastated' and 'appalled' to describe how they felt on discovering the error. I am aware that many affected individuals described the Federation as uncaring and unsympathetic, sometimes expressing negative views about the written and telephone communications they had received. Under the circumstances this is understandable, and there may have been problems with communication style under the stress of the moment, but this is not the general impression I received. Only on one occasion did a Federation respondent propose that medicine is a challenging profession, and that affected candidates might perhaps take it in their stride, a view I wholeheartedly reject. There were many in the Federation who reflected on the duty of candour, which indicated that the error had to be made public by the Federation, despite the negative consequences for the Federation itself. This duty of candour is an important and valuable principle for all healthcare deliverers.

5.4 Technical recommendations

5.4.1 Data Included

One of the signals that was missed was the mean Point BiSerial, which was markedly lower than normal as a result of the presence of 2 apparently correct options for some 80% of the ROP items. This statistic may seem rather arcane, but in fact a Discrimination Index (either PBS or an equivalent) is routinely generated by commercial assessment platforms.

Recommendation 1: The Results Analysis should include the mean PBS or Discrimination Index for each exam, and for preceding papers in the same category.

I note in passing that the mean PBS for the Part 2 Written exams is relatively low, and suggest, short of a recommendation, that this be explored as part of the item review process.

5.4.2 Written Standard Operating Procedures

I heard that not all processes had written Standard Operating Procedures, available to be checked as part of Quality Management Processes, and, importantly, able to guide new staff, or staff standing in for colleagues who were unavailable. The protocol for handling anchor items for SCE exams was specifically mentioned to me.

Recommendation 2: that written Standard Operating Procedures be in place for each step in the exams process, and that these have been followed is checked regularly as part of the Quality Management process.

5.4.3 Manual Processes

A surprising number of processes in the development, delivery and analysis of exams are manual, particularly where there is transition between two different software applications. These are particularly vulnerable to mistakes being made, as was indeed the case with the first error in 23/3. The new Exam Management System is in development, though there are still concerns about its construction, implementation and the engagement of all staff involved in the exams process, as described below in Section 5.2. However, as a technical recommendation, these manual handling steps should be significantly reduced in number.

Recommendation 3: that manual handling steps must be eliminated wherever possible in the development of the Exam Management System.

5.4.4 Exam Results Presentation to the Board

Currently, data is presented largely in the form of small histograms and graphs, whose ongoing statistical significance is hard to gauge. Data tables should be included as more informative, and previous exam performance should be presented as running means and standard deviations, so that outlying results are clearly identifiable.

Recommendation 4: Exam results presentation should be reviewed so that the data is presented in a clear, detailed and comprehensive manner which is self-explanatory for non-expert psychometricians.

5.4.5 Relationship with the Exam Delivery provider

The relationship between the Federation and the delivery provider is plainly crucial but also seemed inadequately codified. This aspect of exam delivery did not seem to be fully in scope for the new Exam Management System.

Recommendation 5: that the relationship with the delivery provider is governed by clearly documented procedures, associated with written Standard Operating Procedures, reviewed regularly as part of the Quality Management process.

Currently, and ostensibly for security reasons, the correct answers are not presented to Surpass, on the grounds that while it would be bad if the paper escaped into the public domain, it would be worse if the answers also escaped. In the era of accessible AI, this is no longer the case: if a paper was released, candidates could readily find the majority of the correct options. The step at which the initial error occurred could therefore be permanently avoided.

Recommendation 6: that the Federation consider providing the correct options to Surpass along with the question papers.

5.4.6 Timing of results release

There was plainly time pressure on the Boards and analysts due to the short (4 week) turn round time. Not only did this mean that the analysis is always carried out under pressure, so too the scrutiny of the results may be rushed. There is of course pressure from candidates to get their results in a timely manner, sometimes exerted on the Federation via the Colleges, but safety must be paramount. There may be time savings emerging in the future in some areas, such as digital recording of PACES results, and these could be well spent on analysis and checking. Of course, the timing of the recruitment rounds must also be borne in mind.

Recommendation 7: that the time between administration and results delivery is sufficient to ensure that there is adequate time for analysis and quality control.

5.4.7 Scaling of Scores

Currently, the original percentage scores are 'scaled' to reflect the mathematics of the standard setting process. The relevant Federation website[22] states that 'the score is higher in candidates who answer more questions correct, but the relationship is not linear and cannot be translated in any meaningful way into a percentage correct score'. But my data indicates that the relationship, while not linear, is monotonic relationship, and that candidates can indeed be given a percentage correct score. In my view this is more useful information than the scaled score, since it tells candidates what percentage of questions they answered correctly overall, just as speciality percentage scores are given to candidates. Of course, it may vary from exam to exam, but I do not see why this common phenomenon should puzzle candidates. In addition, publication of this information may help the Boards scrutinise the results. The pass score did indeed rise slightly under the influence of the 23/3 error, and this could have been yet another warning sign. It would give the Boards an indication of the relative difficulty of each exam, something worth monitoring regularly.

Recommendation 8: that the Federation consider publishing the overall percentage correct score to candidates, along with the scaled score: and the percentage pass mark for that exam, and that the percentage pass mark also be made available to the Boards on an ongoing basis, as part of their monitoring of outcomes.

5.4.8 Should a different standard setting method be used?

The Federation uses test equating (see Section 1.6) in all its written assessments, rather than alternatives such as Angoff and Ebel methods, and concern was expressed to me that this was problematic, particularly with regard to Specialty Certificate Exam, where the pool of candidates is smaller, and marked fluctuations in pass rate occur. It is worth, therefore, briefly reviewing the pros and cons of this approach.

In terms of positives, test equating is probably the most defensible of all standard setting methods in maintaining a consistent standard[23] (although it does not of course guarantee that this is the correct standard). It is also relatively economical in terms of the time required to conduct the procedure, compared to the

extended review of every item required by Angoff and Ebel methods. This is an important consideration given the voluntary nature of the input of many clinicians to the exam process.

However, there are also disadvantages. Since it is a post hoc method, it cannot be commenced until the exam in question has been completed by all candidates, and therefore it adds to the time stresses in releasing the results to candidates. It requires an appropriate number of candidates to sit the exam in order to be robust, and it is not clear what that minimum number is. And since it requires sophisticated analysis using Item Response Theory, it is carried out using the Winsteps programme, which is not readily transparent to clinical members on the Exam Boards who are not psychometricians. As a result, the clinical members of the Boards, and other Federation staff are not involved in standard setting the paper and therefore do not build up the detailed familiarity with the standards that they would do with other standard setting methods. And finally, it may represent a challenge to test security, since it requires items to be re-used from previous exams in considerable numbers. If items have escaped, this will disturb the analysis of the outcomes.

While not recommending the abandonment of test equating, I would suggest that for SCEs in particular, where UK candidate numbers may be quite small, a 'back up' method might be available in case there are very marked variations in the pass rate. One example is the Bookmark standard setting method[24], which would allow clinical members of the Board to use their clinical and educational expertise to apply a standard setting method which might help alleviate their concerns. I do not make a specific recommendation that this method is used, merely that the standard setting methods for SCEs are reviewed with these thoughts in mind.

I also heard other concerns about SCEs, particularly around the data used for anchor items, and it would be a valuable contribution to unified working, if Federation staff and officers, and Board members were to reach a consensus on what was needed in these regards.

Recommendation 9: that standard setting methods for the SCEs are reviewed with a view to reestablishing confidence in both the methods used and the quality of data for anchor items.

5.5 Organisational Recommendations for the Federation and Royal Colleges

5.5.1 Federation Structure

In my view, the primary cause of the two errors in 2023 lay in the relationship at that time between the Federation and the Colleges. The situation has improved since late 2023, with the appointment of a Federation CEO who has an equal voice with the College CEOs. Some processes, such as insurance for the Federation, are still handled by RCP London, and complexities might also arise over, for instance, choices of software where the preferences of the Federation and of a College were not identical. Given the legal, managerial and financial complexities of the situation, it is not possible for me to indicate what a more effective structure would be, and it is outside the scope of my review, but the essence required for secure exam delivery is that the Federation should have sufficient say in strategic decisions to be able to identify and implement strategies necessary to deliver the exams securely in a timely manner. While the Federation also delivers important activities in training and CPD, the risk and cost of failure in the exams process is greatest and should be seen as the highest priority.

Recommendation 10: that the relationship between the Federation and Colleges be structured so that the resources necessary for secure exam delivery are available in a timely manner.

5.5.2 Exam management process

Currently, the written examinations are managed separately with regard to paper creation (the 'Editorial' team) and results management (the 'Research Unit') This seems irrational: these are two ends of the same process and should be integrated into a single exams team. The name "Research Unit" is anomalous, since its most important and demanding role is results handling and analysis – there is a separate Research and Development unit (ironically, with more staff than the 'Research Unit'). If the Research Unit had been involved with paper creation, they might have been more aware that the file sent to Surpass contained the option 'A' as the 'correct' answer to all questions, since this was in the file sent by Federation editorial staff.

Recommendation 11: that the Federation create an integrated team, under clear leadership, responsible for the whole exam process from item writing, through paper development, to results analysis.

5.5.3 Results Review

The process for reviewing results before they are released to candidates is inadequate in relying essentially on the decision of the Board Chair, with advice only from the Head of the Research Unit, based on data which is hard to understand. The review process need not involve the whole Board, since this would make it difficult to schedule, but rather an expert sub-committee of the Board, featuring those Board members with particular assessment interests and/or psychometric expertise. Obviously, this review must be scheduled to be as soon as possible after the results are complete, and appropriately placed in the timetables of the individuals involved.

Recommendation 12: Results should be reviewed by the Board Chair and an expert sub-committee of the Board before release to candidates

Even when the data is as clearly presented as possible, it will still require some expert interpretation. A process of induction for Board members, particularly those of the exam review subcommittee, would be invaluable. Currently clinical Board members approve and monitor the clinical content of the exams, but have to take test equating and exam statistics largely on trust. It became clear to me that there were substantial areas where Board members did not fully understand the processes involved in standard setting and results handling, in large measure due to the complex and opaque nature of the standard setting process.

Recommendation 13: Technical Induction for clinical Board members in the generation of results and processing of outcomes.

5.5.4 Face to face meetings

While virtual meetings represent a considerable saving in travel time and resource costs (including the carbon footprint) it was frequently represented to me that the degree of engagement is greater at in person meetings, and that issues are more closely interrogated where physical presence is available. Currently, the Federation relies to a considerable extent on the goodwill of volunteer clinicians, and making meetings more interactive may not only promote good practice and engagement, but also act as a positive incentive to take part in the demanding activities currently required.

Recommendation 14: that the Federation consider making Board meetings hybrid.

5.5.5 Alleviation of time pressures

One of the issues that arose was the time pressure to release results of the 23.3 exams to a specified time scale, which did not allow for sufficient analysis and reflection on the results to take place. Of course, this pressure is in part due to the natural desire of candidates to obtain their results as soon as possible, and certainly in time for the next round of applications for further exam diets or training. However, the secure delivery of exam results must be paramount. The implementation of all-electronic approaches rather than 'paper and pencil' methods which require secure postage and then scanning, may enable time scales to be more relaxed. And if the correct options are indeed signalled to Surpass, another step may be removed from the timetable, easing this process.

Recommendation 15: that the timetable for results release should include adequate time for checking, and guidance to candidates that results release may be delayed for good cause.

5.5.6 Quality Assurance

I believe that there is a need for a dedicated quality assurance process for assessments, to review the quality of management processes and to oversee the implementation of clear written protocols for transition and manual handling points of the data flow. The quality management process should also include review of the Risk Register, to ensure that the risks are adequately reflected, and that the mitigations are actively processed. The assessment process must have integrated quality management, extending from exam development all the way through to the release of results, within this general quality management process.

Recommendation 16: that an Exams Quality Assurance Management Committee be created, that particular attention be paid to integrated quality management of the exam process from beginning to end, and that regular review of the Risk Register, including Mitigations, should be part of this process.

As I understand it, quality assurance is part of the remit of the Management and Policy Board, which receives reports from the Exam Boards. However, this is just one of the responsibilities of the MAP Board, and I believe

it would be better, for such a mission critical issue as exam delivery, if a dedicated body oversaw the exams quality management process, and then reported in a unified manner to the Board.

5.5.7 Moving to the new Exam Management System

I was surprised to hear that the incoming Exam Management System (EMS) seemed more focused on candidate level data than transparent handling and display of exam outcomes. At each end of the exam process, from exam development all the way to results handling, I was not clear how relevant assessment-facing staff were engaged with the EMS development and design, and as a result, I believe there are still risks relating to the integration of the EMS with the assessment processes. Once the EMS is in place, it will be important to ensure that it can be readily interrogated, and provides clear and comprehensible outputs, for instance using Power BI to generate dashboards and visualisations.

Recommendation 17: that the development of the EMS is integrated with the examination processes, enabling clear non-technical data presentation and straightforward interrogation by Boards, staff and officers.

5.5.8 Team training for the new EMS

There will need to be extensive training for staff in the use of the new EMS. My concern is that the relatively small body of staff currently processing the exam data will have this added as an extra burden, either risking their current work, or their full and expert adoption of working with the new system.

Recommendation 18: that training in the use of the new EMS is paralleled with sufficient support to ensure that the current assessment system is operated safely.

5.5.9 PACES delivery

I was informed that the fact that while PACES is developed centrally by the Federation, it is delivered independently by the three Colleges in the UK and by the Federation internationally means that slightly different administrative procedures are employed in each setting, for instance in candidate databases. I believe that it would be safer if the methods used were as similar as possible, for example by using common software and recording approaches, to help ensure both security and the maintenance of a common standard.

Recommendation 19: that the Federation and Colleges consider how to unify PACES delivery and administrative systems to the greatest extent possible.

5.5.10 Moving to digital recording for PACES

Plans are already underway to move to digital scoring via tablets during PACES to eliminate the necessity for pencil and paper records, scanning, and consequent errors introduced by the scanning process. It will be important to ensure that the digital output is seamlessly integrated into the new Eam Management System, and outcomes can be readily visualised and interrogated as part of the review process.

Recommendation 20: that the move to digital scoring for PACES is seamlessly incorporated into the Exam Management System, with clear outputs and easy interrogation of outcomes.

5.5.11 Exam Security

One respondent raised the issue of exam security. Remote Online candidate-venue delivered exams always pose a challenge to security, even with proctoring. Candidates may have a variety of ways of checking answers or recording questions, and no proctoring system can identify all of these. Escape of items is a particular problem with test equating from anchor items, since this requires the anchor items to be repeated across different sittings of the exam. The rapidly developing capability and availability of AI are also a significant challenge to Remote Online Proctored exams.

Recommendation 21: that close attention is paid to developing security risks with regard to ROP examinations.

This activity would best be carried out in association with other medical Royal Colleges.

5.5.12 Succession Planning

Given that the test equating standard setting method currently used is complex, and currently only the small number of 'Research Unit' staff are capable of implementing it, it seems to me essential that there is succession planning for retirement or other staff departures with regard to this process.

Recommendation 22: that succession planning for staff involved in delivering the standard setting method be implemented.

5.6 General Considerations for Professional Testing Bodies

In scope for my review was to identify general considerations on errors which might be of value to other postgraduate medical education testing and selection bodies, and indeed I believe that such considerations might extend to other high stakes testing assessments such as the Solicitors Qualifying Examinations.

5.6.1 What do professional exams measure?

One might assume that exams simply measure knowledge. But doctors in training have all graduated from recognised medical schools, have worked in clinical environments, and have access to a wide variety of educational materials. In these circumstances, personal factors become more significant. One of these is cognitive ability, including the ability to form mental schemas which aid in memory recall. A second is conscientiousness, which is a good predictor of later exam performance. Then there are demographic factors. In some assessment settings, age matters, with younger candidates often doing better than older ones. There may be gender differences, with females doing better or less well than males depending on the setting. And, of considerable concern, there are ethnicity differences, observed across a wide range of healthcare education (and other) settings. These ethnicity differences may be quite small in scale at the level of performance (perhaps just a few percentage points) but are amplified by their proximity to the pass mark, which may transform a 2% difference in performance into a 5 or 10% difference in attainment.

Most of these factors are stable for the individual. As a result, candidates tend to perform consistently across the exams they undertake. This observation is relevant (see Section 4.3) to the understanding of how signals which accompanied the error which brought about this review were set aside after the exam.

There is a significant body of evidence that connects low exam performance to higher subsequent likelihood of fitness to practice sanctions[25] and to later clinical performance and patient care outcomes[26]. Of course, this relationship is statistical and general: it does not mean that all candidates who pass an exam are safe, and all those who fail are hazardous. And much depends on the environment in which doctors practice subsequent to the examination.

5.6.2 Should specialty written exams be replaced by other assessment methodologies?

In discussions of the various alternative methods for remediation proposed by candidates, such as consideration of ARCP outcomes, portfolios, MultiSource Feedback, or other approaches (see Section 3.7), the concern was raised by Royal College and Federation representatives that adopting alternative methods on this occasion would weaken the case for having professional exams at all. I do not believe this is the case. Under routine conditions, while they are not a magic panacea, written exams are an acceptably valid, reliable and cost-effective means of assessing capabilities, and may also have positive educational impact, in that they stimulate intensive study and preparation on the part of candidates. While approaches such as programmatic assessment[27] are currently widely promoted in medical education, time must elapse for them to gather the body of predictive validity evidence already available for written exams in medical contexts. Very Short Answer questions may offer some advantages, but also have disadvantages, both with regard to standard setting and to the lack of predictive validity evidence[28]. My proposal that alternative assessment methods should be considered in the event of an exam error being discovered within a certain time frame after the results have been released (Section 5.7.4) should not in any way be taken as a recommendation that written (largely MCQ) exams as currently employed by Royal Colleges and other testing bodies should be done away with. As indicated in the previous section, there is a substantial body of evidence that indicates that exam scores are indeed associated with good clinical practice in the long run.

There is, however, a concern over exam security, either by improper means being used during an exam, or by items being released in advance of it. Ongoing consideration must be given to how such high-stakes exams are delivered, with Remote Online Proctoring just one of the options, despite how economical it is to deliver.

5.6.3 Should there be a 'Statute of Limitations' for a future examination error of this kind?

A number of respondents, including Federation officers and other senior doctors, raised the question of what would happen if an error of this kind had been discovered, say, 10 years after the administration of an exam – if for instance an error in medical school graduation exams was subsequently discovered. As the current Part 2 error shows, affected individuals may well have moved on in their lives and careers to the point where just resitting the exams is neither feasible nor appropriate. Under such circumstances, alternative responses should be available, and some possibilities are discussed below. And of course, the length of time which must elapse before such alternative methods are invoked is a matter for speculation rather than calculation, and any suggestion will by its nature seem arbitrary. However, I believe it is at least worth making the attempt here, if only to offer a starting point for future discussions. It appears that the decisions currently made in the case of the Part 2 error are already fixed: nonetheless, I hope that the general argument has merit. I have considered the scenarios of a result discovered (a) within one year after the erroneous results were released, (b) between one and three years after the results were released and (c) longer than three years after the results were released. I focus on candidates who were erroneously told they had passed when in fact they had failed. A year is not simply a calendar year: depending on when major recruitment or other assessment events take place it may be slightly longer or shorter.

(a) If such an error is discovered within a one-year cycle, I believe it is appropriate, though challenging, for candidates to be required to pass the same exam, after, of course, the refund of exam fees, the expunging of the attempt from their record, and support for study in the form of materials etc.
(b) Between one and three years, I believe candidates should be required to show they have reached the required standard, but that alternative methods by which this is established should be employed. For instance, components of the assessment could be offered at suitable intervals rather than large single exam diets on defined dates. This strategy is expanded in section 5.7.4.
(c) After three years, it seems to me unrealistic and unjust to require those affected to sit exams from which they are now likely to be distant. Instead, an appropriate way of ensuring patient safety could be to conduct the equivalent of a Revalidation review, focussing on patient safety issues, and including evidence of safe practice, continuing professional development and ARCP outcomes if relevant. Should any challenges short of referral to the Fitness to Procedures procedure have arisen, these should be addressed by remedial retraining.

5.6.4 Alternative methods of assessment in the event of delayed discovery of an error

In the event of delayed discovery of an exam error, an alternative to requiring candidates to resit the same assessment in the standard format irrespective of their particular circumstances, would be to consider splitting the exam into components, which could be more readily undertaken by affected candidates in their very different circumstances from a conventional candidate. A strategy could be imagined in which these components were tested by individualised methods such as Linear On the Fly Testing (LOFT), or Computer Adaptive Testing (CAT). Such tests could be delivered, uniquely to each candidate and asynchronously, through the use of Item Response Theory, at any time, massively reducing the temporal burden on candidates of preparing for fixed diets. It is true that LOFT and CAD require extensive item banks. However, the advent of AI and its power to generate new items, if carefully curated, means that this requirement becomes achievable, for the first time. Associated with distance learning packages produced by the examining body, who after all were at fault here, of the kind suggested by some affected candidates, these could be undertaken by candidates at a time of their own choosing, making it much easier to reconcile their circumstances with the requirement to demonstrate the correct standard. It is essential to preserve standards: but the standard is not uniquely defined by one style of exam.

It is important to note that I am not suggesting this style as a replacement for all higher exams. As I indicate in Section 5.7.1 there is positive educational impact in preparing for the standard exam. But candidates affected by an error have already met the requirement to study and prepare for the standard exam, perhaps several times. Then they were disadvantaged by an error which was not their fault. They would not be resitting a standard exam under equitable conditions. The Gold Guide[29] to Postgraduate Training allows for different approaches to be taken in the event of *Force Majeure* such as pandemic or "due to cancellation or postponement of a required examination" and discovery of an exam error after a significant passage of time might be considered in the same light.

5.6.5 Should there be an 'External Examiner' for high stakes exams?

One suggestion made to me was that there may be merit in having an external person involved in the exam process within Colleges, perhaps someone with assessment engagement in another College. While the External Examiner system in UK Medical Schools is far from perfect, it does provide the possibility of independent oversight by someone not encultured into the ways of a particular programme, in the way that the Federation processes suffered from an element of complacency as described in Section 4.1.2.

5.6.6 Assessment Error Review

In a medical setting, if a serious adverse event had happened, there would almost certainly be something like a Morbidity and Mortality Conference, in order to identify what happened and where practice can be improved. This would typically include a wide range healthcare staff. If such a meeting had been held with regard to previous errors within the Federation exam processes, such as the PACES error in 2022, with the specific remit of asking "where in our assessment processes are there other vulnerabilities and risks?" then I believe it is at least possible that lessons would have been learned that reduced the likelihood of a another error occurring again. I would propose that any Royal College (or other high stakes professional assessment body) that discovers an error in its processes, *even if it was detected before release,* should hold such an open-forum, no-blame meeting with staff at all levels to explore where else might be vulnerable. If there was an 'external examiner' present at such a meeting, this might help guard against institutional complacency.

5.6.7 The role of the GMC

The GMC's statutory duty is to maintain (a) patient safety (b) standards and (c) the well being of the affected doctors, in that order. Unfortunately, these principles are to an extent at odds with each other. Doctors were so severely affected by the requirement to resit the exams in their standard form despite the lapse of time, that they themselves became patients. In turn, their existing and future patients were affected, with doctors not merely going off sick, having sleepless night, or being distracted while at work, but also by the disruption to career planning and advancement.

I completely concur that standards must be maintained, both with regard to knowledge and skills exams, but where a substantial period of time has elapsed before an error is discovered, I believe alternative methods of re-assessment should be considered, by which the required standard is demonstrated. And finally, the decision taken itself posed serious challenges to the wellbeing of doctors, as this review indicates, and it is not clear to me that this received the fullest possible consideration, even though it was raised in discussion.

One topic that was discussed with the GMC at an early stage was the idea that candidates within one standard error of measurement below the pass mark might be allowed to pass nonetheless (what is known as a condoned pass). This idea was rejected, correctly in my view, since this would have been a clear indication that the standard had not in fact been reached.

The GMC indicated that they can use their convening power to bring together organisation to explore ways forward, and described how they had attempted to promote doctor wellbeing by engaging with various stakeholders such Trusts, Local Education Providers and other training frameworks.

5.6.8 The Role of the Statutory Education Boards

The decision by the SEBs to withdraw applicants for HST from the interview process was also one which was made rather rapidly, and perhaps without full consideration of all the possibilities. I have heard that there was legal representation from both Scotland and England at the relevant meeting, and that an issue was expressed about candidates who had sat the exam on other occasions and who might have a legal case if it transpired that they had perhaps been displaced by a candidate who had received an erroneous result. It might have been valuable to establish the parameters, the numbers involved, and other possible remedies for such candidates before coming to a decision.

5.6.9 Taking time to respond to future errors

It is clear that rapid decisions about how to proceed after the error was discovered were made by the Federation, the GMC and the SEBs. This is understandable, since the information was leaked and began to circulate quite early on, and those affected would be extremely anxious to know what the implications would be for their future practice. But it also meant that various alternatives, of the kind suggested in Section 5.7.4 could not be developed and considered. Nor could the wide range of circumstances in which candidates found themselves be taken into consideration. As a result, the conclusion was very much the conventional approach

of saying that the exams would have to be undertaken again in their standard form, without full knowledge of how this might impact on candidates. Should such a circumstance, of erroneous results being released followed by delayed discovery that this had taken place, arise again, it might be better to have a more extended approach, in which the range of impacts on candidates, and various alternative solutions, could be considered, with greater input from those affected and their representatives. Perhaps this review might offer evidence of the possible impacts of examination errors, which could be taken into account on future occasions. This might allow decisions to be made speedily, but also with insight into possible consequences.

Section 6 How this Review was conducted

6.1 Methodology

A short time and intensive timescale of three months was proposed and acted upon, in order that candidates should receive the fullest possible information in as timely a manner as possible.

6.1.1 Interviews and Written Submissions

A dedicated and secure e-mail account was created for the purpose of confidential communication, and invitations circulated via the Federation to all affected candidates and a variety of other stakeholders. Virtual interviews were usually carried out via Teams or Zoom. 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. A semi-structured approach to the interview was employed, in that there were themes that might be explored, but essentially the process was driven by the respondents and their preferences. Respondents were asked to confirm into which category they fell (e.g. having been passed when they had failed, or vice versa, and at what stage in their career they were then and now). Data saturation was reached, at which point no new themes were emerging, towards the end of the review process.

All interviews were conducted with guarantees of confidentiality, but the possibility of anonymised and non-attributable quotes in the final Review was indicated. Interviews were generally scheduled for 30-60 minutes, though more time was generally available if required. I made myself available over a flexible 'early morning to late evening and weekends' schedule, to allow doctors working shifts or internationally the best possible chance to contribute.

A number of affected doctors submitted written accounts describing their situations and the consequences of the error for them. These written accounts were unprompted and spontaneous, though in a few instances, clarification was subsequently sought.

Overall, I heard from some 50 affected candidates, covering all general categories of those impacted by the error. I also had access to anonymised phone logs of conversations with almost 100 candidates, although of course these may have duplicated some of my interviews and candidate communications.

Interviews with key individuals from the Federation assessment processes were also conducted, with an initial contact list of key individuals being provided by the Federation. These included senior officers of the Federation and Colleges, administrative staff, examiners, members of the Part 1, Part 2 and SCE Boards and representatives of candidate organisations such as the Resident Doctors Associations.

Invitations were also extended to the GMC, with whom a valuable meeting was held, and to the Medical Directors of the Statutory Education Boards, who did not respond.

An open invitation was extended to other stakeholders to contact me directly, and a number of individuals made use of this opportunity. Direct invitations were issued to individuals and organisations to contribute, and in the end, some 40 individuals were spoken with, including some representing larger groups of individuals such as the Resident Doctors Committees.

Fortnightly update meetings were arranged with the Federation Chief Executive Officer, Rachael O'Flynn and Dr Hany Eteiba, President of the RCPSG. These allowed urgent actions proposed by the reviewer to be fed into the planning cycle of the Federation as soon as possible, and plans developed by the Federation during a separate Internal Review to receive external oversight and comment. As a result, a number of suggestions and recommendations which I made were integrated into the Federation's practices and policies during the time period of the Review. Of the formal Recommendations I make, a number therefore are already in train.

6.1.2 Document Review

A range of documents were proactively supplied by the Federation. I requested others, such as Minutes of meetings, correspondence, and internal documents relating to the errors. I requested a variety of data files, relating to the 23/3 and other exams. All such requests were met by the Federation in a timely manner. Documents informed both the interviews, and the final Review document.

6.1.3 Rapid Literature Review

A rapid literature review was conducted around various themes, including the predictive validity of exams, and technical psychometric issues relating to standard setting. Social media and press documentation was also explored.

6.1.4 Thanks and acknowledgements

I would like to record my thanks to the respondents from the Federation and Colleges, who provided me with frank and honest accounts of what had happened, openly acknowledging individual and collective errors in the exam processes. Most of all, however, I would like to express my gratitude to those affected doctors who engaged with me, for sharing with me honest, painful and distressing details of how they felt that they had been, and were being, affected by the errors, often with high degrees of emotion. I hope that being heard was of value to them, even if I could not include all their comments to me in this review or meet all their wishes in terms of changes to decisions already made before my review started. I would like to offer them every possible best wish as they continue to navigate the challenges they are facing.

6.1.5 About the Author

The author is currently Professor of Medical Education and formerly Interim Head of School at UCLan Medical School. He has been 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), and by the Department of Health (as academic partner reviewing revalidation for doctors), amongst others. In 2023 and 2023, he carried out reviews of exam processes for the Royal College of Anaesthetists and the Royal College of Emergency Medicine. Currently he is Psychometric Advisor to the Recruitment Development Group of the UK Foundation Programme Office. Previously he was a Board Member of the UK Clinical Aptitude Test, and Editor-in-Chief of *Medical Education*, the leading journal in the field. He has published widely on assessment in health care settings. In 2022, he was awarded the Gold Medal of the Association for the Study of Medical Education for his services to the field.

Section 7 References

[1] https://pmc.ncbi.nlm.nih.gov/articles/PMC8514562/

[2] Knowledge Based Assessment
[3] IM_Curriculum_Sept2519.pdf
[4] Format | The Federation
[5] https://www.thefederation.uk/examinations/specialty-certificate-examinations
[6] https://risr.global/
[7] https://surpass.com/en-gb/
[8] https://www.winsteps.com/index.htm
[9] De Champlain, A.F., 2018. Standard setting methods in medical education: high-stakes assessment. *Understanding medical education: Evidence, theory, and practice*, pp.347-359.
[10] McManus, I.C., Chis, L., Fox, R., Waller, D. and Tang, P., 2014. Implementing statistical equating for MRCP (UK) Parts 1 and 2. *BMC Medical Education*, *14*, pp.1-19.
[11] https://www.thefederation.uk/news dates from 19th February 2025 onwards
[12] https://doi.org/10.1136/bmj.r534
[13] https://www.independent.co.uk/news/uk/home-news/bma-doctors-exams-passed-failed-test-b2701904.html

[14] https://uk.news.yahoo.com/calls-compensation-hundreds-doctors-received-135800823.html

[15] https://www.dailymail.co.uk/health/article-14419549/Fury-blunder-lets-medics-stay-work-failing-exam.html

[16] https://inews.co.uk/news/doctors-failed-exams-specialist-nhs-results-blunder-3543153

[17] Wiegmann, Douglas A. PhD*; Wood, Laura J. MS*; Cohen, Tara N. PhD†; Shappell, Scott A. PhD‡. Understanding the "Swiss Cheese Model" and Its Application to Patient Safety. Journal of Patient Safety 18(2):p 119-123, March 2022. | DOI: 10.1097/PTS.0000000000000810

[18] https://www.england.nhs.uk/wp-content/uploads/2020/11/Revised-Never-Events-policy-and-framework-FINAL.pdf

[19] Mimende, S., Van Gog, T., Van Den Berge, K., Van Sause, J.L. and Schmidt, H.G., 2014. Why do doctors make mistakes? A study of the role of salient distracting clinical features. *Academic Medicine*, *89*(1), pp.114-120.

[20] Crockery, P., 2003. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, *78*(8), pp.775-780.

[21] https://en.wikipedia.org/wiki/Hadiza_Bawa-Garba_case

[22] Equating overview.pdf

[23] McManus, I.C., Chis, L., Fox, R., Waller, D. and Tang, P., 2014. Implementing statistical equating for MRCP (UK) Parts 1 and 2. *BMC Medical Education*, *14*, pp.1-19.

[24] Lewis, D.M., Mitzel, H.C., Mercado, R.L. and Schulz, E.M., 2012. The bookmark standard setting procedure. In *Setting performance standards* (pp. 225-253). Routledge.

[25] This section, while not a full review of the topic, lists a number of relevant sources of evidence.

Tamblyn, R., Abrahamowicz, M., Dauphinee, D., et al., 2007. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA, 298, pp.993–1001.

Papadakis, M.A., Arnold, G.K., Blank, L.L., Holmboe, E.S. and Lipner, R.S., 2008. Performance during internal medicine residency training and subsequent disciplinary action by state licensing boards. *Annals of Internal Medicine*, *148*(11), pp.869-876.

Tiffin, P.A., Paton, L.W., Mwandigha, L.M., McLachlan, J.C. and Illing, J., 2017. Predicting fitness to practise events in international medical graduates who registered as UK doctors via the Professional and Linguistic Assessments Board (PLAB) system: a national cohort study. BMC Medicine, 15, pp.1–15.

Cuddy, M.M., Young, A., Gelman, A., et al., 2017. Exploring the relationships between USMLE performance and disciplinary action in practice: a validity study of score inferences from a licensure examination. Academic Medicine, 92, pp.1780–1785.

Arnhart, K.L., Cuddy, M.M., Johnson, D., et al., 2021. Multiple United States Medical Licensing Examination attempts and the estimated risk of disciplinary actions among graduates of U.S. and Canadian medical schools. Academic Medicine, 96, pp.1319–1323. https://doi.org/10.1097/ACM.0000000000004210

Cuddy, M.M., Liu, C., Ouyang, W., et al., 2022. An examination of the associations among USMLE Step 3 scores and the likelihood of disciplinary action in practice. Academic Medicine, 97, pp.1504–1510.

Bartman, I., Kain, N., Ashworth, N., Hernandez-Ceron, N., Hurava, I., Hamayeli-Mehrabani, H., Kumar, K., Nie, R. and Morin, M., 2024. Do Canadian Medical Licensing Exam Scores Correlate with Physicians' Future Performance in Practice? A Cohort Study of Alberta Family Physicians. *Journal of Medical Regulation*, *110*(4), pp.13-19.

[26] This section, while not a full review of the topic, lists a number of relevant sources of evidence.

Ramsey, P.G., Carline, J.D., Inui, T.S., Larson, E.B., LoGerfo, J.P. and Wenrich, M.D., 1989. Predictive validity of certification by the American Board of Internal Medicine. Annals of Internal Medicine, 110, pp.719–726.

Tamblyn, R., Abrahamowicz, M., Brailovsky, C., et al., 1998. Association between licensing examination scores and resource use and quality of care in primary care practice. JAMA, 280, pp.989–996.

Tamblyn, R., Abrahamowicz, M., Dauphinee, W.D., Hanley, J.A., Norcin, J., Girard, N., Grand'Maison, P. and Brailovsky, C., 2002. Association between licensure examination scores and practice in primary care. JAMA, 288, pp.3019–3026.

Hamdy, H., Prasad, K., Anderson, M.B., Scherpbier, A., Williams, R., Zwierstra, R. and Cuddihy, H., 2006. BEME systematic review: predictive values of measurements obtained in medical schools and future performance in medical practice. Medical Teacher, 28, pp.103–106.

Tamblyn, R., Abrahamowicz, M., Dauphinee, D., et al., 2007. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA, 298, pp.993–1001.

Norcini, J.J., Boulet, J.R., Opalek, A., et al., 2014. The relationship between licensing examination performance and the outcomes of care by international medical school graduates. Academic Medicine, 89, pp.1157–1162.

Tiffin, P.A., Illing, J., Kasim, A.S. and McLachlan, J.C., 2014. Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: national data linkage study. BMJ, 348.

Sharma, A., Schauer, D.P., Kelleher, M., et al., 2019. USMLE Step 2 CK: best predictor of multimodal performance in an internal medicine residency. Journal of Graduate Medical Education, 11, pp.412–419.

Hamstra, S.J., Cuddy, M.M., Jurich, D., et al., 2021. Exploring the association between USMLE scores and ACGME milestone ratings: a validity study using national data from emergency medicine. Academic Medicine, 96, pp.1324–1331.

Asemu, Y.M., Yigzaw, T., Desta, F.A., Scheele, F. and van den Akker, T., 2024. Does higher performance in a national licensing examination predict better quality of care? A longitudinal observational study of Ethiopian anesthetists. *BMC anesthesiology*, *24*(1), p.188.

Norcini, J., Grabovsky, I., Barone, M.A., Anderson, M.B., Pandian, R.S. and Mechaber, A.J., 2024. The associations between United States Medical Licensing Examination performance and outcomes of patient care. Academic Medicine, 99(3), pp.325–330.

Gray, B.M., Vandergrift, J.L., Stevens, J.P., Lipner, R.S., McDonald, F.S. and Landon, B.E., 2024. Associations of internal medicine residency milestone ratings and certification examination scores with patient outcomes. *JAMA*, *332*(4), pp.300-309.

[27] van der Vleuten, C., Lindemann, I. and Schmidt, L., 2018. Programmatic assessment: the process, rationale and evidence for modern evaluation approaches in medical education. *Medical Journal of Australia*, *209*(9), pp.386-388.

[28] Potter, H.G. and McLachlan, J.C., 2025. Assessing medical knowledge: A 3-year comparative study of very short answer vs. multiple choice questions. *Medical Teacher*, pp.1-9.

[29] https://www.copmed.org.uk/images/docs/goldguide10thedition/Gold%20Guide%2010th%20Edition%20August%202024.pdf. See “Outcome 10” Page 79.