The Council for European Specialist Medical Assessment (CESMA – UEMS)

Cork, Ireland- Friday May 2nd and Saturday May 3rd, 2025

Day 1 Friday May 2, 2025 , 13:00 – 18:00

Chairs: Maeve Durkan, Ambrogio Fassina, Danny Mathysen, Albert Mifsud, Gian Battista Parigi

1pm – 2pm

Chair Maeve Durkan

Dr. Maeve Durkan welcomed all attendees to the CESMA Spring Meeting in Cork and expressed her appreciation for their participation. She introduced the CESMA Executive Committee and then each delegates were invited to introduce themselves briefly by passing the microphone around the room. Each attendee shared their name, affiliation, and area of interest.

Spoken Remarks by Dr. Maeve Durkan, CESMA Spring Meeting, Cork

We are observing the ongoing development of the European Board Examinations. Europe needs standardized assessments due to diverse national training systems.The exams offer a benchmark for consistency and quality across countries.Our goal is to create valid, fair, and reliable assessments. We must ensure only competent candidates are certified. This meeting allows us to share experiences and address key challenges.We need to manage marking variations and candidate inconsistencies.Fairness, transparency, and quality assurance remain central.

Online and off-site exams raise concerns about system reliability and security. We must ensure oversight, fairness, and strong exam infrastructure. Decisions are needed on written, oral, and OSCE formats, and MCQ structure. AI offers innovation but raises concerns over integrity, privacy, and bias.These topics will be explored further in tomorrow's session.

Prof. Danny Mathysen – Presentation Overview - "All You Wanted to Know, but Didn't Dare to Ask: Designing High-Quality MCQs"

Multiple-choice questions (MCQs) are often perceived as easy to create and inherently objective. They are scalable, well-suited for online delivery, and widely used across Europe. However, in practice, developing high-quality MCQs requires substantial time, expertise, and rigorous review processes. A key metric in item analysis is the facility index ; the percentage of candidates who answer a question correctly. The optimal range for this value is typically between 15% and 85%. This range helps ensure that items effectively discriminate between low-, mid-, and high-performing candidates and allows for visual mapping of item performance across ability levels.

It is also important to recognize that psychometric properties are not static; they vary depending on the candidate population and the overall composition of the question pool. A question that performs well in one exam cohort may underperform in another, emphasizing the need for continual item review and contextual interpretation of psychometric data.

When constructing multiple-choice questions (MCQs), it is best to use a clinical scenario format consisting of three short paragraphs: the first describes the patient's presentation (including age, setting, symptoms, and history), the second outlines a few key investigation findings, and the third poses a clear, direct question ending with a question mark.

Answer options should be consistent in length and structure, and should avoid giving internal clues, contradictions, or including overly detailed distractors. Common mistakes to avoid include using negatively worded questions, revealing answers in other questions, and offering choices such as "all of the above" or "none of the above."

Review Process and Quality Control

Regular peer review and blueprint alignment ensure MCQs match the syllabus and ETR. Post-exam changes should be avoided to maintain content balance. Creating high-quality questions takes about an hour each, making training and structured review essential. As Prof. Mathysen noted, good assessments rely on planning, collaboration, and rigorous review — not just statistics.

Second Best Answer Scoring

When post-exam analysis reveals a second defensible answer, full marks may be awarded to both options. To preserve fairness and transparency, such scoring policies should be pre-declared where possible and these decisions must return to the question review panel for long-term item adjustments.

Negative Marking Policy

The group confirmed a move away from negative marking, citing concerns over its limited discriminatory validity, the risk of unfairly penalizing cautious candidates, and its disproportionate impact on non-native English speakers. Positive scoring systems are now preferred, particularly in high-stakes licensing contexts, where fairness and candidate equity are critical.

Exam Blueprinting and Content Mapping

Exam design should follow a structured blueprinting process that aligns with defined learning objectives and distinguishing between critical and optional content. Question types should be mapped according to the required cognitive level (e.g., recall, understanding, application). Postgraduate exams typically include 100–130 high-quality, well-constructed items.

Multiple Answer Questions

Multiple Answer Questions, where more than one response may be correct, are increasingly used for their resistance to AI-solving. These formats are perceived as more difficult but offer enhanced reliability and require clear candidate instructions and careful validation during item writing.

Timing Guidelines for SBA Questions

Standard time allocation is 90 seconds per question with a clinical scenario and 60 seconds per question without a clinical scenario. These timings are inclusive of candidates with disabilities and account for multiple reviews of a question within the exam window and extending time beyond this may not improve performance and can encourage candidates to revisit questions unnecessarily.

Determining the Number of Questions

The number of exam questions should be determined by blueprinting analysis, taking into account the importance of topics and the frequency and severity of conditions in clinical practice. Arbitrary targets should be avoided unless justified by the blueprint. Both over-testing (which causes redundancy without improving reliability) and under-testing (which reduces scope and reliability) should be avoided. Generally, 100–130 questions are adequate for high-stakes European exams, while some specialties, such as cardiology, may use 100–180 questions depending on exam complexity and candidate numbers. Smaller institutional exams can use proportionally fewer questions.

Effect of Exam Length and Extra Time

There is no strong evidence that extending exam time substantially improves candidate scores. In fact, extended time may increase fatigue, promote second-guessing, and heighten the risk of item exposure in loosely proctored environments. While it is essential to ensure time fairness for candidates with approved accommodations, applying a universal time extension is not always beneficial and may even be counterproductive

2pm -2.15 pm

Challenges to Setting Up an Exam – Dr. Kaushik Chaudhuri (2:00–2:15 PM)

A European-level specialist examination project, jointly led by UEMS and EULAR, has been developed to create a European certification exam ensuring consistent, high-quality specialist training. Earlier efforts were delayed by lack of consensus, but the initiative was revived two years ago in response to the growing need for standardized training and cross-border recognition of medical qualifications.

The project aims to enhance training quality, align with European scientific society goals, and build a sustainable examination model. Major challenges have included gaining stakeholder agreement, harmonizing diverse national systems, ensuring psychometric fairness, developing a secure digital platform, and maintaining a team of skilled question writers and reviewers.

A task force of 20 contributors has produced a bank of about 200 questions (180 finalized), introduced standard setting based on proven models, and validated the exam structure and blueprint, drawing inspiration from the UK and Psychiatry Examination Board experiences. The first 90-question exam is planned for November 2025, accompanied by training workshops for item writers and reviewers. The exam will operate under independent validation and governance, maintaining academic integrity while recognizing partner contributions. The fee is expected to be €600–700, with clear candidate information, sample questions, and FAQs provided in advance. Ultimately, the project demonstrates that developing a European specialist exam is a complex, collaborative, and resource-intensive process, but one that strengthens unity, quality, and sustainability in European medical training—laying the foundation for a respected qualification for future generations.

2.15 -2.30 pm

Intro – Maeve Durkan/ Prof Mathysen

Eligibility and Logistics – Maeve Durkan / Prof. Mathysen (2:15–2:30 PM)

Eligibility and Candidate Verification

The panel discussed the challenges of defining and verifying eligibility for European specialist exams, including inconsistent national standards and difficulties confirming credentials, especially from regions with weak documentation. Candidates usually need 3.5 years of training and two references, verified through professional trust networks. Growing interest from non-EU applicants highlights the need to balance inclusivity with maintaining European certification standards.

Points for Future Consideration

The debate centered on whether European exams should stay limited to trainees within Europe or include international applicants. While openness promotes global recognition, it raises verification and quality concerns. A centralized digital verification system under UEMS could help manage this. Exam boards must uphold European standards and ensure ethical, transparent governance.

3.00 – 3.30 PM Coffee break

3.30 – 4 pm / 4-5pm

Chair Maeve Durkan
Q WRITING / Best of 5 – Presenter RCPI Keith Farrington 15 mins
How many Questions do you need ? Keith Farrington/ Prof Danny Matthysen –
Discussion Q& A 10 mins
Psychometrics – What to look for ? Discussion Q & A. 5 minutes

What Makes a Good Question?

A well-designed MCQ is built around a realistic clinical scenario that includes relevant history, examination, and investigation details. It should be followed by a clear lead-in question and five options, one correct answer and four plausible distractors. The goal is to assess clinical reasoning and knowledge application, not to trick candidates.

Question Purpose: Discrimination Over Difficulty

The most effective questions go beyond testing factual knowledge, they distinguish between strong and weak candidates by assessing clinical judgment and reasoning. A well-written question should be answered correctly by high-performing candidates, be sufficiently challenging to discourage random guessing, and include plausible distractors that require thoughtful consideration from less-prepared candidates. The goal is not to create trivia or trick questions, but to evaluate decision-making in the context of safe, evidence-based patient care.

Writing Pitfalls to Avoid- Common problems include

Avoid common MCQ pitfalls such as leading questions, negative phrasing, implausible distractors, and overly simple cases that fail to test deeper reasoning. All answer options should be consistent in format, grammar, length, and logic to maintain fairness and clarity. MCQs should pass three stages: (1) peer review by 2–3 reviewers, (2) a technical check for formatting and clarity, and (3) committee feedback with possible rewriting.

What Makes a Question Difficult?

Difficulty doesn't always mean complexity. Often, it comes down to the quality of the distractors, whether the candidate has clinical context or experience and how long it takes to process the information presented. A question that requires interpreting investigations, applying guidelines, or thinking through several steps of reasoning is inherently more difficult and often more useful for assessment purposes.

Avoid Negative Impact on Candidates

High-stakes exams require clarity, fairness, and consistency. Avoid ambiguous wording, unnecessary complexity, and culturally specific content without context to ensure a reliable candidate experience. Question writers must standardize or clearly present measurement units to prevent confusion across countries. Ultimately, questions should be clear, accessible, and focused on assessing clinical competence, recognizing the diverse and often distracted realities of today's learners.

Distractor Quality and Post-Exam Analysis

We recently conducted a two-month post-examination analysis, which revealed important insights into question performance. In one instance, 99% of candidates selected the correct answer, a result that may seem positive at first glance but, in reality, indicates that the question failed to discriminate between levels of ability. When a distractor is *non-functional* that is, when no candidate selects it. It ceases to serve its intended purpose of differentiating weaker candidates from stronger ones. Effective distractors should challenge those with incomplete understanding while remaining clearly incorrect to well-prepared candidates.

Reuse, Validity, and Learning Effects

Some examination questions are reused across multiple years, which may seem efficient but poses significant validity concerns. Over time, repeated exposure to identical or similar items can lead to familiarity effects where candidates perform better not due to increased competence, but because of prior awareness of the question structure or wording. To preserve the integrity and discriminative power of assessments, it is essential to regularly retire, revise, and replace questions, particularly those that have been widely circulated or discussed. Creating high-quality examination questions is both a science and an art. It involves adherence to technical principles such as clear phrasing, functional distractors, and consistent terminology while also requiring professional judgment to ensure that each item is reasonable, relevant, and fair. The overarching goal remains to assess clinical reasoning, uphold patient safety, reward competence, and support meaningful learning.

Q Writing and Psychometrics (3:30–5:00 PM) Chair: Maeve Durkan

- **Presentation: Best of 5 Q Writing – Keith Farrington (RCPI) – 15 mins**
- **Discussion: How Many Questions Do You Need? – Keith Farrington / Prof. Mathysen – 10 min**
- **Discussion: Psychometrics – What to Look For? – 5 mins**

Key performance metrics in exam analysis include the P-value, which indicates item difficulty (the proportion of candidates who answered correctly); the Rit value (item-total correlation), which measures the item's ability to discriminate between high- and low-performing candidates; and the Standard Error of Measurement (SEM), which reflects the precision and reliability of candidate scores. Together, these psychometric indicators provide essential data for item review, validation, and ongoing question bank maintenance, ensuring that each question contributes effectively to a fair, reliable, and valid assessment.

Q Writing Workshop (4:00–5:00 PM) Facilitator: Keith Farrington (RCPI)

The workshop focused on an interactive peer review of Best-of-Five questions submitted anonymously by delegates. Delegates were asked to bring one question for group discussion.

Key areas of feedback included:

- Clarity and conciseness of question stems
- Appropriateness and plausibility of distractors
- Relevance to the curriculum and blueprint

The session also addressed common challenges such as the use of abbreviations and potential sources of confusion. Delegates were encouraged to ask questions and provide comments throughout, with a strong emphasis on collaborative improvement and adherence to best practices in question writing.

4.00pm-5.00pm.

Q WRITING WORKSHOP Facilitator RCPI Keith Farrington Best of 5
Delegates to bring ONE Q each (anonymous) and will be reviewed at workshop

Workshop Overview

The session opened with Keith Farrington, who stressed that high-quality question writing and review are crucial for creating fair, valid, and reliable exams.

- Best-of-Five (Bo5) Format: Each item presents a clinical scenario, a clear lead-in, and five options, only one correct, testing applied clinical knowledge.
- Role of Distractors: Strong, plausible distractors help differentiate candidate ability; weak ones undermine validity.
- Discrimination: Effective questions should clearly separate well-prepared candidates from weaker ones, as illustrated through shared examples.
- Exam Blueprint: A comprehensive blueprint ensures alignment with curriculum goals and balanced topic representation.
- Style and Clarity: Questions must be clear, concise, and free of ambiguity; trick questions and overly complex stems should be avoided. A consistent style guide supports uniformity.
- Review Process: Reviewing typically takes under five minutes per item, emphasizing the need for well-prepared initial drafts despite large item volumes.
- Question Difficulty: Driven mainly by candidate expertise and distractor quality, with additional influence from topic complexity and stem length.

Overall, the session underscored that thoughtful structure, clarity, and rigorous review are the foundation of effective assessment design.

5.00 pm – 5.30pm

Chair Maeve Durkan
Standard setting / Lead RCPI Keith Farrington. 10 minutes

Standard Setting Overview and Workshop
Chair: Maeve Durkan / Lead: Keith Farrington (RCPI)

Why Standard Setting Matters? Who is the Borderline Candidate?

The discussion emphasized that the purpose of the examination must be clearly defined before designing or setting standards. Whether the exam targets early trainees, those completing specialty training, or serves for certification, progression, or feedback, its intent determines the appropriate level of difficulty and standard setting approach.

Standards should not be set for top or unprepared candidates but for the "borderline" or minimally competent candidate—someone just meeting the required standard. To define this level, examiners must ask:

- What does this candidate know?
- What can they do?
- Would I trust them to manage patients under supervision?

This concept guides fair, realistic standard setting, ensuring that exams reflect the competence required for safe clinical practice.

Standard Setting in Practice

The exam uses a criterion-referenced approach, meaning candidates are judged against a fixed performance standard, not against each other. The Modified Angoff method is the preferred standard-setting technique. Each of 5–9 trained judges independently estimates the percentage of borderline candidates likely to answer each question correctly. Judges review questions without knowing the correct answers, considering clinical context, difficulty, and chance factors. After individual scoring, the panel reconvenes to discuss discrepancies and refine estimates. This collective review ensures consistency and credibility in determining the final pass mark.

Establishing the Pass Mark

The average of judges' ratings from the Modified Angoff process determines the recommended pass mark. Question difficulty depends on several factors: the complexity of reasoning, how common the topic is in practice, the clarity of options, and local variations in clinical norms. Items can range from very easy (85–100%), moderately difficult (40–60%), to very difficult (0–20%) for borderline candidates. Examiners should ask whether each item is fair, discriminates effectively, and reflects real-world competence expected of a just-passing trainee. Though standard setting can seem technical, it is a structured, collaborative, and vital process that ensures fairness, safety, and academic integrity in determining readiness for specialist practice.

Workshop (10 mins)

The session on standard setting in practice was led by Keith Farrington and Prof. Danny Mathysen, with contributions from Dr. Kaushik and Dr. Morne. Discussions focused on managing low pass rates, the lifespan of exam questions, and maintaining robust judging panels. Keith highlighted that exam items should not be reused after three years to preserve validity and relevance, stressing the need to continually expand the question bank. The group also noted the difficulty of recruiting and retaining seven trained judges for standard setting, emphasizing the importance of consistency, experience, and expertise within these panels to ensure fairness and quality across exam cycles.

5.30-5-50pm

Dr Morne Wolmarans

Oral Examinations
Presenter: Dr. Morne Wolmarans

Oral exam / OSCE – 10 minutes. How to set up – 1O mins
How to score 5 minutes
Costs and Logistics numbers – 5 minutes

Many candidates show strong theoretical knowledge but limited clinical application, prompting the introduction of the OSCE to assess practical skills objectively and consistently. The OSCE complements written exams by improving validity, reliability, and fairness. Each OSCE session lasts 10 minutes per candidate, with 10 minutes for setup and brief periods for scoring and logistics explanation. A typical OSCE includes up to six stations, each with two questions and two rotation cycles to cover a broad range of competencies. Assessment combines point-based scoring with rating scales such as fail, borderline, pass, and excellent.

Dr. Wolmarans emphasized objectivity as a core principle, confirming that the structured format ensures fairness and consistency. Candidates receive all necessary information online before the exam, which is included in the exam fee. For large cohorts, candidates arrive early, mobile phones are prohibited, and separation prevents sharing of questions. Online exams use randomized questions to maintain integrity and avoid repetition. Examiners must complete training before assessments, with logistics including travel, accommodation, catering, and scheduling.
Keith Farrington and others discussed the importance of balancing feasibility, cost, and quality when scaling examinations. Participants agreed that continuous improvement and examiner training are vital to ensure consistency, fairness, and high standards across all assessments.

Dinner 7.00 pm Orchids Restaurant – Hayfield Manor

Day 2

8.30 – 9.15

Update Treasurer – Prof Parigi 15 minutes

Dr. Gianni Battista Parigi presented an overview of CESMA's current financial status, noting that income from scientific sections remains stable, with several showing positive growth. Major expenses this year included meeting organization and a €6,000 allocation for executive travel in Brussels. Bank fees remain low, and payment processes are handled efficiently.

Between 2023 and 2024, CESMA incurred €2,200 in management fees, consistent with the typical annual range of €1,000–€1,500. The organization's investment in U.S. bonds, totaling €5,800, continues to perform well and remains accessible as a financial reserve. After challenges in 2020 due to the pandemic, CESMA's finances have recovered steadily, moving from deficit to a sustainable surplus. The 2024 surplus largely stems from resumed appraisals and successful event organization. CESMA generally holds two two-day meetings per year, costing around €7,500 annually, which is considered reasonable for their impact.

In his final report as Treasurer, Dr. Parigi expressed gratitude for the committee's support and reaffirmed that all meeting funds have been received and finances are in good order. He concluded by thanking colleagues for their trust and cooperation during his tenure.

Update President – Dr Maeve Durkan 5 minutes

A Congress will be held in Leuven in 2026, and preparatory work is already underway.
Further details about the programme, logistics, and participation will be announced in due course. Because of the extensive planning required and our active involvement in organizing the event, no spring meeting will take place this year. All collective efforts will focus on supporting Congress preparations and ensuring strong representation of the Section. Members' continued support, collaboration, and engagement are greatly appreciated as we work toward a successful Congress.

Update appraisals – Prof Danny Mathysen 15 minutes

Several sections have requested examination feedback, and coordination continues to improve communication and clarity. Collaboration among societies is increasing, but alignment with European standards remains essential. The goal is to strengthen partnerships, promote knowledge exchange, and uphold exam integrity. Regional variations persist, but compliance with European frameworks and openness to global cooperation are priorities.New models are being explored to enhance participation and accountability.The next system meeting will be held in December at the Brussels office to review the first appraisals.Applications should be submitted six months before exams, or at least three months in advance.The appraisal process takes about six months, from question writing to final review.Guidance documents, application forms, and report templates will be available on the website.Preparations for the 2027 examination cycle begin in early 2026, emphasizing timely submissions and efficient coordination.

9.15 – 10.00

As part of the agenda, the election for the new CESMA Treasurer took place. All nominated candidates were given the opportunity to briefly introduce themselves prior to the start of voting. Candidates were invited to present in alphabetical order, with each allotted 3–5 minutes to introduce themselves, outline their experience and motivation, and share their vision for the Treasurer role. The following candidates presented themselves in the order listed:

1. Dr. Vesna Kushec
2. Dr. Kaushic Chaudery
3. Dr. Alexandru Nica
4. Dr. Morne

Following the introductions, the voting was carried out accordingly.

10.00 – 10.45

Recap Friday workshop

Recap what happen on Friday and then open the floor for questions

Steps in exam development (6)
Q number – Blueprint / Major areas/ minor areas / 'weighting Qs'

Q writing – Good quality Q / 1 hour / 3 paragraphs – cognitive load / 90 secs/ No abbreviations/ Definitive Q
Talk one hour to write a good question and important not to have a distractor.
Q reviewing – Reformatting / '5 minute rule/ 1 set units/
Q Standard setting – allow minimum 10% above Q on paper / Q number / Cut score ±2SD/ METHODOLOGY ?
Q Final review
Q Post exam* analysis

Oral exam vs OSCE and Logistics/ costs/ benefits

Exam Development Workshop

Crafting clear, relevant questions is essential to effective exam design and must align with the Training Volume or curriculum to reflect core competencies. Blueprinting defines the exam's key knowledge areas and ensures balanced coverage across specialties for fairness and consistency. All questions undergo systematic review for clarity, bias, and alignment, involving subject experts and psychometric advisors. Exam delivery requires careful planning of logistics, security, and format (online or in-person) to maintain integrity. Post-exam analysis includes item performance review, statistical evaluation, and standard setting to establish defensible pass marks. The process must remain transparent, structured, and evidence-based, involving collaboration among experts, appraisers, and technical teams.

Blueprinting should mirror real-world clinical relevance and ensure proportional weighting based on condition frequency and severity. Underperforming questions should be reviewed and blueprinting refined over time to improve validity and balance. A dedicated workshop covered syllabus development, blueprinting, question weighting, and psychometrics to support credible exam design. Participants were reminded that understanding psychometrics enhances transparency and confidence, even without doing statistical calculations. Future masterclasses should target both appraisers and exam designers, combining theoretical and practical sessions. Suggested formats include modular sessions with live case studies held during major CESMA meetings. The group strongly supported developing this training to strengthen credibility, transparency, and quality in European assessments. Next steps include proposing dates, selecting expert facilitators, and deciding whether the event will be online, hybrid, or in-person.

Meeting Summary – Part 2: Question Writing & Assessment Design

Writing high-quality MCQs requires approximately one hour per question, especially to design meaningful distractors. Question stems should be carefully structured and avoid "back-of-envelope" writing to maintain clarity and educational value. Distractors must be plausible and relevant; implausible or "killer" distractors reduce quality and fairness. "Killer" options that contradict guidelines or reflect unsafe care should be avoided, as they frustrate candidates and lack educational value. Some participants argued such options may reflect real-world errors, but consensus favored avoiding them except in rare, justified cases. Average exam timing is around 90 seconds per question, adjustable depending on complexity, with some exams using 75 seconds as a guide. Additional time for candidates with dyslexia or neurodiverse conditions appears fair and does not significantly affect overall outcomes. Formative use of exams is encouraged to identify learning gaps before final certification, supporting targeted trainee development. There is strong interest in future workshops on blueprinting, syllabus alignment, question writing, psychometrics, and review processes. These sessions may be tailored for different roles, such as appraisers, examiners, and curriculum developers.

Open questions remain about separating workshops by role, the appropriateness of killer distractors, ideal timing per question, and accommodations for neurodiverse candidates. Including implausible options risks promoting guessing behavior rather than genuine reasoning and may unfairly penalize careful candidates. Rarely, implausible distractors can serve confirmatory purposes in high-stakes settings but must be justified and transparently used. Reliability improves with question volume: fewer than 60 questions yields poor reliability, 60–100 is borderline, and over 100 is optimal. Longer exams (e.g., 200 questions) should be divided into two blocks to reduce fatigue and maintain focus. Exams must balance factual recall with higher-order thinking, emphasizing reasoning, integration, and clinical decision-making. Different specialties require tailored approaches—surgical exams may focus on anatomy, while medical exams prioritize applied knowledge and management.

RCSI Marie Morris – comments & observations 10 minutes

10.45 – 11.00
Picking/ updating your exam provider -
How and why did you pick who you did
What do you want ? How much have you got ?
The Qs to ask

Dr. Morris opened by thanking participants and commending the professionalism and quality of the meeting presentations.With 18 years of experience in medical education, including at Trinity College Dublin and the Royal College of Surgeons in Ireland, Dr. Morris emphasized extensive expertise in curriculum and assessment design. She acknowledged significant progress in professionalising, standardising, and validating medical assessments across Europe.

Blueprinting was highlighted as vital for guiding assessment focus and ensuring exams measure the most relevant aspects of clinical competence.Dr. Morris stressed balancing assessment of learning with assessment for learning to support trainee development.Recommendations included using multiple assessors, repeated evaluations, and mock or peer assessments to enhance learning.The importance of validity and constructive alignment was emphasized to ensure learning outcomes, teaching, and assessment are coherently linked.Dr. Morris warned that fragmented curricula undermine educational coherence and fairness.For postgraduate training, assessments should evolve from knowledge recall toward demonstrations of applied skills through OSCEs and workplace-based evaluations. Blueprinting was described as essential for fairness, transparency, and validity across subspecialties and exam formats. Attention was drawn to human factors such as examiner fatigue, which can threaten consistency and fairness in evaluations.Spreading assessments over multiple days and supporting examiner wellbeing were recommended as quality assurance measures.

In closing, Dr. Morris praised the progress in European medical and surgical education and urged continued commitment to fairness, transparency, validity, and patient safety.
The address concluded that effective assessments should not only measure competence but also enhance learning, professional practice, and patient care outcomes.

11 – 11.20 Coffee break

11.20 – 11.50
Q : Prof Ashraf Butt – 10 minutes
How are you awarding your exam – Discussion
- diploma
- Fellow

Who is eligible to take your exam ?

Skewing of pass mark – How to overcome that ?
'Skewed ' cut scores

Diplomas, Fellowships, and Medical Certification Standards: A Reflection

This section explores the role and meaning of diplomas and fellowships in medical education from professional and public perspectives. It notes the difficulty of comparing international qualifications, as systems vary in emphasis on exams, training structure, and pass standards. From the public view, a diploma signifies study completion, but passing an exam alone does not prove full clinical competence, it only reflects performance at one point in time.Across Europe, diplomas focus on specific skills, while fellowships represent advanced training, specialization, and continued professional engagement. Fellowships embody mentorship, contribution, and belonging within a professional community. These differences complicate international equivalence, as some systems link certification to structured experience, while others rely mainly on exam results.Such variation challenges fairness, comparability, and public confidence.CESMA plays a central role in promoting harmonization and consultant-level standards across Europe. Ultimately, true competence arises from supervised, experiential training, and diplomas or fellowships should represent both achievement and sustained capability to preserve trust and integrity in medical certification.

Uniformity in Fellowship and Diploma Recognition Across Regions

This section emphasizes the need for clarity, fairness, and consistency in medical certification, credentialing, and recognition. While fellowship is seen as a prestigious mark of excellence, disparities persist in how equivalent qualifications are granted across countries.
Examples such as Poland and Argentina reveal inequities, candidates sitting the same standardized exam may receive different recognition, with Europeans earning a fellowship and non-Europeans only a certificate. Such discrepancies undermine the credibility and fairness of a supposedly universal examination system.
Trainee feedback shows that fellowships hold greater professional value than diplomas or certificates, symbolizing peer recognition and achievement. This raises a key equity question: if candidates meet the same standards and performance levels, why should recognition differ?
Eligibility rules add complexity, European candidates must be in accredited training, while non-European applicants may qualify through experience and references. This inclusivity increases access but highlights the need for consistent recognition when identical standards are applied.
Finally, the section reaffirms that exam success reflects performance at one moment, not lifelong competence, which depends on continued development, supervision, and evaluation.

Core Themes from the Discussion on Fellowship vs Diploma and Medical Training Recognition

It was observed that examinations involving soft skills, such as communication, body language, and scenario management—can be rehearsed, meaning candidates may initially perform well in OSCEs but later revert to their natural behaviors. This raises concerns about the objectivity and validity of examiner-based assessments, as observed performance may not accurately represent real clinical practice. Furthermore, examiner bias and fatigue, especially toward the end of the assessment day, were acknowledged as genuine risks that can affect consistency and fairness in candidate evaluation.

Inconsistency in Credential Recognition (Europe vs non-Europe)

It was highlighted that candidates from both EU and non-EU countries often sit the same board examinations and demonstrate comparable performance. However, a discrepancy persists in post-exam recognition: European candidates are typically awarded fellowships, while non-European candidates receive only a certificate, despite meeting identical standards. This inconsistency undermines the principle of fairness and risks diminishing the perceived value of the certification process. Consequently, there were strong calls for uniformity in credentialing, encapsulated in the guiding principle of *"same exam, same recognition."*

Fellowship vs Diploma – Structural and Legal Implications

Fellowships were described as advanced post-specialist training with structured curricula, governance, and mentorship, while diplomas represent shorter, skill-focused programs without full specialist scope. Under EU law, only medicine, nursing, and dentistry are formally recognized as sectoral professions, raising concerns about using "diploma" without ECTS or training duration alignment. Participants noted ongoing challenges in validating non-European training, stressing that identical exam performance should yield equal certification regardless of geography. A single-tier system, potentially adopting the unified title "Fellowship," was proposed to simplify processes, support workforce needs, and enhance fairness, though retrospective reclassification of certificates was seen as administratively difficult.

11.50 – 12.30

AI assisted MCQ writing – Q & A
AI and pass score – Q & A

- **Pass / cut scores and 'skewed data'**

AI in medicine
Prof Danny Mathysen / 15 minutes – Q & A 5 minutes

The discussion explored the role of artificial intelligence (AI) in exam development, recognizing its efficiency benefits while emphasizing significant ethical, security, and quality concerns. AI can help draft MCQs for well-defined topics but often produces hallucinated or implausible content, weakening authenticity and educational value. Risks include data leakage, academic misuse, and cheating, especially if AI systems access or replicate real test materials.

Studies show that GPT-based models can sometimes match or surpass human performance, raising issues of fairness and exam security.AI cannot assess clinical reasoning, judgment, or interpersonal skills, so it should remain a supportive tool, not a replacement for human examiners. Institutions were advised to use closed, secure AI systems and implement strong quality assurance to protect data and maintain exam validity. User feedback showed that only 20% of AI-generated questions were high-quality, while most required substantial editing, reinforcing the need for expert oversight. Concerns about content leakage through open systems like ChatGPT highlighted the need for restricted, institutionally managed platforms. Participants also warned of blueprint drift, where AI favors common topics and neglects balance across domains, stressing that humans must ensure curriculum alignment. The consensus was that exams should not simply become harder in response to AI but should shift toward assessing higher-order reasoning and ethical decision-making.
Finally, multiple-answer questions were noted to challenge both AI and humans, underscoring the importance of clear design and balanced difficulty.

12.40 – 12.55 am
How do we craft the narrative (including points from December meeting)
The December 2024 meeting marked a key step in advancing fairness, consistency, and collaboration in European specialist assessments.Progress was noted in exam development, standard setting, and blueprinting.AI's supportive but limited role in question writing was discussed, emphasizing ethical safeguards.

12.55 – 1pm
Presentation of appraisal awards.

The presentation of appraisal awards has been postponed to December 2025 to align with the updated scheduling and preparation timeline.

1pm– 1.15
What to do when the exam goes wrong ?
Examples of what has gone wrong -
RCPUK problem 2024 with markings for Part 2 _ Speaker TBC 5 minutes / discussion 5

Delegates engaged in a constructive group discussion focused on improving examination quality, harmonizing certification standards, and strengthening collaboration across European sections.

Training Diversity & Exam Necessity

There is considerable variation in specialty training across Europe, shaped by factors such as geography, program design, institutional capacity, and local healthcare priorities. For example, trainee exposure to rare endocrine disorders can vary widely between countries and even among training centres. In this context, examinations must be designed to reflect these differences, maintaining standards that are both equitable and relevant across diverse healthcare systems. Achieving this balance requires close collaboration among educational institutions, specialist societies, and clinical experts to ensure that assessments remain aligned with the realities of medical training and clinical practice throughout Europe.

Our Collaborative Approach

The UEMS Section and Board of Endocrinology works in close partnership with the European Society of Endocrinology and, since 2022, with the Royal College of Physicians of Ireland, which acts as its professional service provider. Together, these organizations aim to identify the minimally competent candidate, a trainee who can apply core scientific knowledge to clinical scenarios, demonstrate critical reasoning, manage both acute and non-acute cases, and recognize red flags in patient care. The scenario-based examination is designed for trainees in their penultimate year of specialist training, ensuring alignment between assessment content, clinical competence, and preparedness for independent practice.

Designing the Exam

The examination blueprint was developed based on the UEMS–ESE curriculum, a landmark publication in the *European Journal of Endocrinology*, ensuring a fair and content-valid assessment. The exam initially comprised 200 questions across two papers, though this has now been refined to 180 questions to improve balance and manageability.

Developing the question bank has been an intensive, collaborative process. Writers received training and refresher sessions, participated in in-person workshops, and submitted questions online, with additional contributions from invited subject-matter experts. Establishing a comprehensive, high-quality bank from scratch required significant time and coordinated effort.

A robust quality assurance framework underpins the process, involving three stages:

1. Editorial vetting of every question,
2. Blueprint alignment checks, and
3. Final committee selection of exam items.

This structured review process ensures that the examination covers all key areas of endocrinology, with particular emphasis on diabetes, reflecting its broad clinical importance across practice settings.

Standard Setting: Angoff Method

The pass mark for the examination was established using the Angoff method, a widely recognized standard-setting process in which expert judges estimate the proportion of borderline candidates expected to answer each question correctly. While effective, the method is inherently judgment-based and therefore subject to some degree of variability or bias, not through error but as a natural consequence of human interpretation. To reduce these effects, several safeguards were implemented: selection of experienced subject-matter experts, in-person moderation sessions to resolve outlier judgments, and monitoring for "hawks" and "doves"—those who consistently score too harshly or too leniently—to ensure balanced and credible outcomes. Exam reliability has also been closely monitored. The Cronbach's alpha improved from approximately 0.7 in 2023 (indicating acceptable internal consistency) to 0.9 in 2024, representing an excellent level of reliability and strong confidence in the exam's measurement quality.

Candidate Factors

Then we asked a different question: Who are our candidates?

- Some were not formally in endocrinology training.
- Others may have been junior internal medicine doctors testing the waters.
- Some appeared to be collecting certificates.
- English proficiency and knowledge of European guidelines varied.
- And, notably, many candidates were from outside Europe.

Hofstee Method: A Compromise

The team enhanced the standard-setting process by combining the Angoff and Hofstee methods to ensure fairness, credibility, and defensibility. The Hofstee approach sets upper and lower limits for acceptable scores and failure rates, balancing rigor with equity. Although only two exam cycles have been completed, this dual-method strategy has proven useful for cross-validating pass marks and increasing confidence in reliability and transparency. Next steps include introducing marker questions to calibrate difficulty, reassessing whether European and non-European candidates should be analyzed separately, and further improving question quality, judge training, and candidate verification.He emphasized building a robust system from the outset rather than adjusting pass marks retrospectively, warning that leniency risks undermining integrity and public trust. Assessment standards must remain defensible, transparent, and educationally sound to safeguard credibility. The session ended with appreciation for participants' engagement and a shared commitment to developing a fair and high-quality certification process for endocrinology across Europe and beyond.

Post-Speech Discussion: Pass Mark, Candidate Heterogeneity, and Data Collection

Maeve reflected on the year-to-year variation in pass rates, noting that despite fluctuations, a higher Cronbach's alpha this year indicated improved exam reliability. Differences in pass rates likely reflect variations in candidate preparedness rather than flaws in exam design or standard setting.
The discussion stressed that if candidates from accredited programs fail, it raises questions about whether minimum competencies for safe practice are being met. Calls to raise the pass mark were rejected to preserve fairness and integrity, as adjusting results would be indefensible and compromise standards of care.

The Board aims to balance participation and excellence, focusing on verified competence rather than pass rates alone. Current certificates confirm exam completion but do not equate to formal qualification, since some candidates are still in training. Greater attention will be given to candidate demographics and training backgrounds to contextualize performance data. Planned measures include updating application forms, collecting more detailed training information, and requiring verified references. The Board reaffirmed its commitment to a fair, defensible, and high-quality examination system that strengthens the credibility and long-term value of European specialist certification.

1.15– 1.30
Reminders of ' legal insertions' – Highlights from Maître Hanoteau 10 minutes
What's your ' back up plan ? 5 minutes
What do you do with the candidate who challenges his/her result ?
Prof Danny Mattysen – ' The call you dread '

Importance of Clear Contracts with Exam Providers

Maître Hanoteau highlighted a crucial legal consideration regarding contracts with examination platform providers. He emphasized the need for explicitly defined responsibilities within these agreements, clearly stating where the provider's obligations end and where those of the examination board or organizing institution begin. Such clarity is essential to avoid disputes, particularly in cases of technical failure or platform collapse, which have occurred in past exam sessions. The key recommendation was that all contracts should include comprehensive liability clauses, ensuring that the exam board or organizing body is legally and operationally protected if the provider fails to deliver the contracted services. This approach strengthens both risk management and the integrity of the examination process.

Candidate Technical Responsibilities

Some technical failures during online exams stem from candidate-side issues, not platform faults. Common causes include poor Wi-Fi, unstable devices, and firewall restrictions, which are not the provider's responsibility. The candidate agreement must explicitly state these responsibilities and include detailed technical guidance.Key requirements include using a wired LAN connection, keeping the device plugged in, avoiding mobile or tablet use, and ensuring firewalls don't block platform access. All candidates must complete a mandatory trial session 1–2 weeks before the exam using the same device, network, and location intended for exam day. Practicing under different conditions often leads to login issues, disconnections, or access failures.Testing from the exact exam environment confirms compatibility and minimizes technical risks. Institutional networks and remote systems frequently block exam platforms, so candidates must verify access in advance.To ensure legal and operational protection, candidate instructions should clearly specify technical requirements, liability disclaimers, and environmental expectations.Next steps include embedding these clauses in candidate and provider contracts, enforcing trial session compliance, and expanding guidance to cover all foreseeable risks.

2.20 pm – 2.45 pm
Setting pass mark and the challenges when more fail than pass ?
Nephrology - Speaker TBC – 10 minutes
Endocrinology – Prof Josanne Vassallo – 10 minutes
OP ED - Opinion _ RCPI / Prof Mathysen

Lunch 1.30 – 2.20pm

2.45 pm -3.15PM
On site or on line exams
What is our future
Comments for our OTLG colleague
Comments from the floor

On-Site or Online Exams — What is Our Future?

Before the COVID-19 pandemic, examinations were predominantly conducted on-site in secure, proctored environments such as Pearson VUE testing centers. The pandemic necessitated a rapid shift to remote online formats, which enabled continuity but also introduced new challenges. As technology continues to advance, examination boards now face a key strategic question: Should future exams remain online, return to in-person settings, or adopt a hybrid approach? The discussion focused heavily on the security and integrity of online exams. Concerns about cheating were acknowledged as real and increasing, with documented cases of inadequate proctoring, use of mobile devices, and poor or absent supervision during remote assessments. It was noted that proctoring providers vary widely in quality, with some offering limited oversight despite extensive marketing claims.

To safeguard exam credibility, it was recommended that boards perform thorough due diligence when selecting or reviewing proctoring services. Key questions should include:

- What specific monitoring mechanisms are in place?
- Can organizers observe live sessions or review recordings after the exam?
- Does the system incorporate AI-assisted surveillance to detect suspicious behavior?

Such measures are essential to ensuring the fairness, reliability, and defensibility of online examinations, and to maintaining public and professional confidence in the integrity of the ***certification process.***

Hybrid & AI-Augmented Solutions

Some examination boards have adopted AI-supported platforms to enhance integrity and monitoring during online assessments. These systems use facial recognition, detect multiple participants, and flag suspicious behavior such as frequent eye movement away from the screen. Real-time proctoring has been implemented with a recommended ratio of one proctor per 20 candidates for effective supervision, while higher ratios were found inadequate. A chief proctor oversees the session and can act immediately on flagged incidents. The platforms feature automated alerts, warning messages, and structured escalation protocols to ensure that all potential breaches are swiftly managed, maintaining fairness and credibility throughout the examination.

Appraisal Experience with Online Proctoring and Limitations of Online Exams

Internal testing of multiple proctoring platforms revealed major differences in reliability and usability, with some appraisers unable to log in and others facing long setup times. These inconsistencies risk undermining candidate confidence, exam integrity, and fairness. It was acknowledged that AI systems can reduce but not eliminate cheating, as determined candidates can still bypass eye-tracking, audio, or room-scanning safeguards. The group discussed the cost–security trade-off, noting that on-site exams offer greater security but higher expense, while online formats are more accessible but require investment in robust systems.Ethically, most candidates act with integrity, yet verification through clear protocols, training, and provider contracts remains essential. Future assessments are expected to adopt a hybrid model, using online exams with remote proctoring for accessibility and on-site exams for high-stakes testing. Everyone stressed that trust must be supported by verification, with technology complemented by transparent processes, accountability, and continuous quality assurance.Ultimately, maintaining fairness, validity, and public confidence requires balancing security, cost, accessibility, and candidate experience.

What's Your Backup Plan When the Exam Fails?

This section addresses the critical issue of examination failure management, emphasizing that such incidents are real and carry serious consequences. Recent cases included a complete platform crash that erased candidate data mid-assessment, representing a total breakdown of the evaluation process. Labeling such events as force majeure offers little reassurance; the real question is when, not if, a failure will occur and whether a contingency plan exists. In one case, having a predefined plan allowed the exam to be rerun within 48 hours, preventing the loss of an entire exam year.

Following these incidents, the board mandated that all providers maintain a secondary server or backup system capable of immediate activation in case of failure.Verification must include details such as data center separation, mid-exam data protection protocols, and tested failover procedures under real conditions. Operational responses differ depending on whether the exam has started, as questions cannot be reused once exposed. Partial data losses, such as incomplete question sets or missing options, were deemed invalid and indefensible.Boards must ensure contractual precision, simulate failures during internal testing, and integrate these checks into quality assurance cycles. As remote examinations grow more common, technical disruptions are inevitable, making provider accountability and system resilience essential.Ultimately, examination bodies are responsible not only for fair assessment but also for ensuring the platform's reliability so candidates can demonstrate their competence without system failure.

3.15 – 3.45pm

Discuss review of exam appraisal – anonymous – proposal to present each' exam appraisal' as a formal discussion for learning

Experiences from appraisers –
Maeve Durkan / Ashraf Butt/ Morne Wolmarans/ Ileana Ghiordanescu

The appraisal process is fundamental to ensuring integrity, transparency, and accountability in examination systems. It verifies that every stage,from development to delivery and review,is documented, defensible, and compliant with established standards. Appraisals provide assurance that, if challenged, the organizing body can demonstrate full procedural alignment and quality control. Although initially intimidating, appraisal is intended as a collaborative learning experience for both appraisers and appraisees, promoting continuous improvement rather than criticism.
A first-time appraiser shared insights from observing a European specialist examination, noting that effective planning, clear deadlines, and responsive organizational support are key to a well-coordinated and credible exam process.

Logistics & Initial Impressions

During the committee meeting, a few minor logistical challenges arose due to the size of the venue and the presence of multiple concurrent scientific events, which made navigation somewhat difficult. Nevertheless, the appraisers arrived on time and engaged in productive discussions with the committee chair, secretary, and reviewers. The opportunity to observe the final round of question review proved particularly valuable. The committee operated efficiently, first dividing into subgroups to review half of the questions in parallel and then reconvening as a full group to finalize and approve the examination content. This process provided a clear demonstration of collaborative decision-making and quality assurance in exam preparation.

Strengths Observed and Concerns Raised

The observations confirmed a robust question review process characterized by collaboration, curriculum alignment, and professionalism.Review teams ensured clinical accuracy, clarity, and consistency, avoiding absolutes, redundancy, and implausible distractors.Discussions were transparent and improvement-focused, refining question design through clearer data presentation, uniform styles, and reduced ambiguity—enhancing fairness and validity.During the online exam, technical issues arose including login failures, latency, and unclear proctor communication, raising concerns about data privacy and transparency.Simulated cheating tests revealed weak detection and escalation mechanisms, showing that real-time monitoring and response were unreliable.
Key recommendations include obtaining explicit consent for remote access, ensuring clear and visible proctor interaction, strengthening screenshot and browser detection, improving incident logging, and setting strict SLAs for proctor conduct, privacy, and system performance.

Standard Setting Issues

Another important observation was that the society did not apply a standardized, evidence-based method for setting the examination pass mark. Instead, the threshold was determined subjectively each year by the organizers. This practice represents a significant area for improvement, as the absence of a formal, scientific standard-setting process—such as the Angoff or Hofstee methods, may compromise the fairness, transparency, and defensibility of the examination outcomes.

Conclusion – A Valuable Experience

Despite initial challenges, the appraisal experience proved highly valuable and educational, allowing the appraiser to apply principles of assessment quality, fairness, and standardization.
Initial apprehension gave way to a collaborative and constructive process, supported by experienced colleagues through reflection and feedback.The experience was described as professionally rewarding and personally motivating, emphasizing the importance of teamwork and continuous learning in medical education.The appraiser expressed gratitude for the opportunity and encouraged wider participation in future appraisal activities.

3.45 – 3.55
Updates on appraisals

4 pm – AOB

The Chair concluded the session on a positive note, thanking participants for their thoughtful contributions, engagement, and collegial spirit throughout the meeting. The past day and a half were described as highly productive, marked by constructive discussions, collaboration, and shared learning that will continue to strengthen collective efforts in the months ahead.

Practical reminders were shared:

- A walking tour of the university will take place the following morning at 10:00 a.m., meeting at the main entrance.
- The December meeting will be held on 6–7 December, avoiding the 13th due to its proximity to the Christmas period.
- The May meeting will be integrated into the Grand MS meeting, with details to follow.
- Future meetings will continue to alternate between member institutions and Brussels, maintaining the balance between engagement and tradition.

In closing, the Chair expressed sincere appreciation to all attendees and wished everyone a pleasant evening and a safe journey home.