An Exploration of the feasibility of establishing a benchmark European examination in Clinical Microbiology

A thesis submitted to the University of Dublin

in fulfilment of the requirements for the degree of

Doctor in Medicine (M.D.)

2024

Dr Maeve Mairead Doyle

MB MRCSI MSc FRCPath FFPath PGDipClinEd

Supervisor: Professor Breida Boyle

Second supervisor: Dr Jane Holland

Declaration, online access and the General Data Protection Regulation

I declare that this report has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's Open Access institutional repository or to allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for 2024.

Student name: Dr Maeve Mairead Doyle

Student number: 0022696

Signature: 

Declaration and Statement of Plagiarism

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Dr Maeve Mairead Doyle

Acknowledgements

The work was performed by the author and undertaken as a member of the European Examination in Medical Microbiology Working Group of The European Union of Medical Specialists (*Union Européenne des Médecins Spécialistes – UEMS*) UEMS, section MM.

I wish to express my deep and sincere gratitude to my supervisor, Professor Breida Boyle for her unwavering support, patience, and advice during my research. I would also like to thank Prof Boyle for her support as a member of the Exam Working Group and secretary of the UEMS section of Medical Microbiology.

I would like to thank Prof Hilpi Rautelin for encouraging my participation in the UEMS section of Medical Microbiology (MM), my involvement in the exam development and my proposal to take on the role of educationalist in the exam working group.

I would like to acknowledge the support and contribution of the members of the European Exam Working Group. I would like to thank Prof Truls Leegaard, the first chair of the working group and now President of the section of Medical Microbiology, for encouraging me to undertake this project. Prof Leegaard wrote the first European curriculum in MM which I converted to an Outcome-Based Education (OBE) curriculum, to create the exam blueprint. I would like to thank Dr Frank van Tiel, current chair of the exam working group, for his support, encouragement and meticulous excel spreadsheets created to check the question-bank progress. I would like to thank Dr Albert Mifsud, a member of the Exam Working Group for his work in helping the group to choose the exam delivery platform.

I would like to thank Dr Jane Holland, for agreeing to be a second supervisor and providing the *real* medical educationalist support for the exam, for reading and reviewing the thesis and associated posters and papers, for delivering question-writing training and then

training me to deliver subsequent training to the question-writers. I would also like to thank Jane for delivering training to the working group in the Angoff method of standard-setting and as a friend, I would like to thank her for encouraging me to continue with the project, despite the global Covid 19 pandemic.

I would like to thank Prof Richard Arnett for undertaking the excellent psychometric analysis which supported the project, for his patience in explaining the results and their implications during the exam delivery and again while I was writing the thesis. I would also like to thank him for guiding us through the Bookmark method of standard setting.

I would like to thank Prof Cees van Der Vleuten for his personal communication and advice regarding the two standard setting methods we used in the pilot exam.

I would like to thank the European Board of Radiology (EBR) and European Diploma in Radiology (EDiR) for the use of their question database and exam delivery platform which we used to deliver the pilot exam in MM, and first real exam in MM, particularly to Vanessa and Violeta for the training they provided and the support during the live invigilation of the pilot exam.

I would like to acknowledge Dr Caoimhe Brennan for her help with the trainees' survey and I would like to acknowledge Dr Grace Chan, Consultant Microbiologist, University Hospital Waterford and Mr Ciaran Carr, RCPI for their help with organising the data from the scope of practice surveys so the result could be analysed.

I would like to thank Mary Murphy, my work colleague and friend, for her help, support, her superior skills in 'Microsoft Word' and her ability to find my misplaced drafts of documents during the project.

I would like to acknowledge both Xenia and Maud, in UEMS, for the administrative support for the Exam Working Group and particularly for the weekly meetings with Dr Frank van Tiel and me in the months leading up to the exam and I would like to thank all the question-writers and reviewers for their work on developing the content for the exam and my colleagues in the UEMS section of MM for participating in the surveys on training and practice.

I would like to thank my mother, family and friends, for their encouragement and particularly my partner John, whose patience, encouragement, kindness and support is endless.

I would like to dedicate the thesis to my father, John Doyle, who sadly passed away on the 6th of July 2023, in loving memory.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Explanation
ADePT	A Process for Decision-making after Pilot and Feasibility trials
AMS	Antimicrobial Stewardship
BDL	Bookmark Difficulty Location
CAP	Child and Adolescent Psychiatry
CbD	Case-based Discussion
CPE	Carbapenemase Producing Enterobacterales
CIT	Combined Infection Training
CM	Clinical Microbiology
CSCST	Certificate for Successful Completion of Specialist Training
CTT	Classical Test Theory
DOPs	Directly Observed Procedure
EBR	European Board of Radiology
ECTS	European Credit Transfer and Accumulation System
EDiR	European Diploma in Radiology
EHEA	European Higher Education Area
EMQ	Extended Matching Questions
ETR	European Training Requirements
EU	European Union
EEMM	European Exam in Medical Microbiology
FRCPath	Fellowship of the Royal College of Pathologists
HSE	Health Service Executive
HST	Higher Specialist Training
ICU	Intensive Care Unit
IPC	Infection Prevention and Control
ISCM	Irish Society of Clinical Microbiologists
IRT	Item Response Theory
KR	Kuder Richardson
LO	Learning Outcome

MCQ	Multiple Choice Questions
MDR	Multidrug Resistant
MDR TB	Multidrug Resistant Tuberculosis
MDT	Multidisciplinary Team Meetings
MEQ	Modified Essay Question
MiniCEX	Mini Clinical Evaluative Exercise
MM	Medical Microbiology
MRC	Medical Research Council
MNSQ	Mean Square Statistics
NDTP	National Doctors Training and Planning
OBE	Outcome Based Education
OSCE	Objective Structured Clinical Examination
OSPE	Objective Structured Practical Examination
PCR	Polymerase Chain Reaction
PLD	Performance Level Descriptors
PMP	Patient Management Problems
1-PL	1 Parameter Logistic Model
2-PL	2 Parameter Logistic Model
RCPI	Royal College of Physicians in Ireland
RP	Response Probability
SB	Spearman Browne
SBA	Single Best Answer
SE	Standard Error
SEM	Standard Error of Measurement
SOP	Standard Operating Procedure
STC	Specialist Training Committee
UEMS	European Union of Medical Specialists (Union Européenne des Médecins Spécialistes)
UK	United Kingdom
US	United States

Abstract

Recruiting doctors in Clinical Microbiology (CM) in Ireland is challenging, and partly hindered by different training & scope of practice in CM across countries. Doctors training in CM in Ireland are required to sit a UK exam to achieve certification of specialist training. Recent changes in that exam have seen a move away from a focus on laboratory skills and Infection Prevention and Control, which are a core part of practice in CM in Ireland. The European Union of Medical Specialists (UEMS) section of Medical Microbiology (MM) created a European curriculum in MM in 2017 to further the UEMS goal of harmonisation of training and practice in Europe and established a group to create a European Exam.

A pilot pan-European Exam in MM was performed in 2021 to simulate each step in the process (curriculum model, question development, blueprinting, standard-setting, reporting results). Candidates completed a questionnaire about their experience of the pilot exam (2021) (Appendix 1), while members of the UEMS section in MM completed two surveys (2019 & 2022) (Appendix 2) to assess progression towards harmonisation of training & scope of practice. Criteria to assess a pilot & feasibility study were applied, based on Thabane's approach and some elements of the complex intervention framework. Ninety-seven candidates from 18 countries sat the 99-item MCQ exam. Rasch analysis indicated a good model fit with optimum distribution of item difficulty & candidate ability. Most candidates found the questions and language clear, and the exam difficulty appropriate, with good assessment of the curriculum. The survey of members of the UEMS section in MM shows increasing harmonisation of specialist training and scope of practice in Europe. As part of the project, a tool for guiding the development of a pilot exam was developed, to facilitate creating exams that are robust, defensible, valid and feasible.

Lay Abstract

Recruiting doctors in Clinical Microbiology (CM) in Ireland is challenging and made more difficult by the different training & scope of practice in CM in different countries. Doctors training in CM in Ireland must sit a specialist UK exam to achieve certification of specialist training. The UK exam has changed and no longer focusses on areas such as laboratory skills and Infection Prevention and Control, which are a core part of the job in CM in Ireland. The European Union of Medical Specialists (UEMS) section of Medical Microbiology (MM), whose goal is harmonisation of training and specialist medical practice in Europe, created a European curriculum in MM in 2017 and established a working group to create a European Exam.

A pilot pan-European Exam in Medical Microbiology was performed in 2021 to simulate each step in the process of developing an exam. Candidates sitting the exam completed a questionnaire about their experience of the pilot exam (2021), while members of the UEMS section in MM completed two surveys (2019 & 2022) to assess how they work and whether there is progression towards harmonisation of training & scope of practice. The pilot exam was evaluated to assess if creating a real European exam in MM was feasible.

Ninety-seven candidates from 18 countries sat the exam. Statistical analysis of the results showed that it is a quality exam and tested the candidate's ability well. Most candidates found the questions and language clear, and felt the exam difficulty was appropriate with good assessment of the CM knowledge. The survey of members of the UEMS section in MM shows increasing harmonisation of specialist training and scope of practice in Europe. A tool for guiding the development of a pilot exam was developed, to help other groups to create exams that are robust, defensible, good quality and feasible.

An Exploration of the feasibility of establishing a benchmark European Examination in Clinical Microbiology

The terms Clinical Microbiology (CM) and Medical Microbiology (MM) are interchangeable, and both will be used throughout the thesis.

Hypotheses of the Project:

In Ireland, successful completion of the Fellowship of the Royal College of Pathology (FRCPath) examination is a requirement for post-graduate medical trainees in Higher Specialist Training (HST) in Clinical Microbiology (CM) to be awarded their Certificate for Successful Completion of Specialist Training (CSCST). Recent changes in the curriculum in Clinical microbiology and Infection in the UK is changing the structure and focus of the FRCPath exam in the UK. Specialists in Clinical Microbiology in Ireland are concerned that future changes in the UK exam may render it less suitable for assessment of trainees in Clinical Microbiology in Ireland.

Hypothesis 1: The development of a European Examination in Clinical Microbiology would give the Irish postgraduate training body for CM an option to explore as an alternative post graduate assessment for the Irish trainees in CM.

The European Union of Medical Specialists (*UEMS: Union Européenne des Médecins Spécialistes*), founded in 1958 is a professional organisation of Medical Specialists in the European Union. The main purposes of the UEMS are the harmonisation and improvement of the quality of medical specialists training, qualifications and practice in the European Union.

Hypothesis 2: The development of a European Examination in Clinical Microbiology would drive harmonisation of medical specialists training, qualifications and practice in the European Union.

Ireland is experiencing a workforce crisis in Clinical Microbiology. There is a difficulty recruiting doctors to fill the current Consultant Microbiologist posts in Ireland. A recent workforce planning review in CM has indicated that there is an urgent need to increase both training and Consultant posts in CM to meet the workload. (HSE and NDTP 2023). The differences in the role of the Specialists in CM in some European countries also makes recruitment of doctors from outside Ireland difficult when trying to fill vacant posts in Ireland.

Hypothesis 3: The introduction of a European Exam in Medical Microbiology (MM) would likely continue to improve and harmonise training and would identify candidates who might be suitably trained to work in MM in Ireland.

Aims:

To explore the feasibility of establishing a benchmark European Examination in Clinical Microbiology

To develop a pilot European Examination in CM underpinned by a robust evidence-based approach to assessment and to simulate each step in the process (the curriculum model, question development, blueprinting, standard-setting, and reporting the results)

To evaluate the effectiveness of a benchmark European Examination in Clinical Microbiology on harmonisation of the training and practice of Microbiology in Europe.

To develop a tool for guiding the development of a pilot exam, to support and guide colleagues creating a new exam to achieve the necessary steps and standards to create an assessment that is robust, defensible, quality-assured and feasible.

Value of Research

It is well established that assessment drives learning (Brown, Bull et al. 1997). The main purposes of the UEMS are the harmonisation and improvement of the quality of medical specialists training, qualifications and practice in the European Union. The introduction of a European Exam in Medical Microbiology would likely continue to improve and harmonise training and practice of CM across Europe. A benchmark European Examination in MM would support this key UEMS goal. This may also facilitate movement of specialists.

MM in Ireland has a workforce and recruitment crisis. MM has an essential role in patient safety, as evidenced by the work undertaken during the global Covid-19 pandemic. The Medical Microbiologists were responsible for leading the introduction and delivery of Covid-19 testing in Ireland. The Medical Microbiologists also lead the Hospital Infection Prevention and Control services, including outbreak management, design of isolation facilities and healthcare ventilation.

The harmonisation of scope of practice in MM in Europe would significantly aid in identifying suitably trained candidates for working within in MM in Ireland.

This project outlines the steps taken to create a quality assured benchmark European Examination, underpinned by the current medical education scientific practices, including curriculum design, blueprinting, question-writing and review and standard setting. The project also considers exam delivery and incorporates feedback from the examinees, using this feedback to improve the examination quality. This work will be useful to other sections in UEMS planning to develop a European Examination. A tool for creating a pilot exam was developed as part of this project.

Outputs

Published Article:

2021 Publication in Clinical Microbiology and Infection. (Impact Factor 13.31 in 2021)

Doyle M, Boyle B, Brennan C, Holland J, van Tiel F, Mifsud A, et al. Specialist training in Medical Microbiology across Europe in 2021 – An update on the actual training situation based on a survey. Clinical Microbiology and Infection. 2021; 27: 1576-80.

(Doyle, Boyle et al. 2021)

Presentations to Learned Societies:

- Doyle M, Boyle B, Holland J, Mifsud A, Van Tiel F, Leegaard TM. 2021. Establishing a Pan-European examination in Medical Microbiology. ECCMID, online.
- Doyle M, Boyle B, Holland J, Mifsud A, Van Tiel F, Leegaard TM. 2022. The trainee experience: candidate feedback from the pilot European Exam in Medical Microbiology. ECCMID 2022. Lisbon, Portugal.
- Doyle M, Leegaard TM, Holland J, Mifsud A, Arnett R, Van Tiel F, Boyle B. 2023. The first European Exam in Medical Microbiology (EEMM): the role of exam candidate feedback in improving the trainees' experience. ECCMID 2023. Copenhagen, Denmark.
- Doyle M, Leegaard TM, Holland J, Mifsud A, Arnett R, Van Tiel F, Boyle B. 2024. Medical Microbiology in Europe: developments in training and scope of practice ECCMID 2024. Barcelona, Spain.
- Doyle M, Leegaard TM, Holland J, Mifsud A, Arnett R, Van Tiel F, Boyle B. 2024. Moving towards harmonisation of clinical practice in Europe: reviewing the development of a

European curriculum and assessment. ePoster – on demand presentation at the International Association for Health Professions Education (AMEE) 2024 Basel, Switzerland.

Chapter 1: Introduction

1.1 Introduction and Clinical Context

1.1.1 Ireland – Clinical Role

Clinical Microbiology is a medical speciality in which the Doctors specialise in the management of human disease caused by micro-organisms including bacteria, viruses, fungi and parasites (Humphreys, Nagy et al. 2010). The role of CM is to advise on the diagnosis, treatment and prevention of infection. CM are experts in laboratory diagnostics for infectious disease, typically the Clinical lead of their Hospital Infection Prevention and Control Service and advise on antimicrobial therapy and management of infection. CM also plays a lead role in Antimicrobial Stewardship, including monitoring and managing antimicrobial resistance. Clinical practice is underpinned by a core knowledge of microbial pathogenesis and pathogen epidemiology. CM interacts with colleagues in all areas of Medicine, Surgery, Paediatrics, Obstetrics and Gynaecology, Occupational Medicine, Public Health Medicine and Community and General Practice. The term Medical Microbiology (MM) may also be used in the above contexts, and the term is interchangeable with the term Clinical Microbiology (CM).

The recent covid-19 pandemic has further highlighted the importance of the role of CM in improving patient safety. In Ireland, the Consultant Microbiologists led on the rapid development and implementation of Covid-19 diagnostics, which has been shown to play a key role in management of the Covid-19 pandemic (Caruana, Patel et al. 2020, Peeling, Heymann et al. 2022). In addition, their leadership role in hospital prevention and control of infection, drawing on their expertise in outbreak management and hospital design,

including ventilation standards, was essential to minimise the transmission of covid-19 to service users and staff (Battey, Doran et al. 2023) (Chandra, Kusum et al. 2022).

1.1.2 Ireland – Training

In Ireland, Clinical Microbiologists are trained on a structured training programme delivered by the Faculty of Pathology, Royal College of Physicians in Ireland. Doctors applying for a place on the training programme in Clinical Microbiology must have successfully completed a post-graduate examination e.g. Membership of the Royal College of Physicians of Ireland (MRCPI), or equivalent, prior to application. In addition to completing the five-year higher specialist programme, the post-graduate medical trainees in Higher Specialist Training (HST) in Clinical Microbiology must also successfully complete a specialist examination in their Pathology discipline. Currently, the exam that is recognised by the Irish HST in Clinical Microbiology is the Fellowship of the Royal College of Pathology (FRCPath), UK examination. Completion of the FRCPath (UK) in Clinical Microbiology is a requirement for trainees to be awarded their Certificate of Successful Completion of Specialist Training (CSCST).

In April 2014, the General Medical Council in the UK approved a new curriculum for training their doctors to become Infection Specialists. In August 2015, the new curriculum was introduced for UK Medical Microbiology, Medical Virology, Infectious Diseases and Tropical Medicine trainees. The new curriculum was a Combined Infection Training (CIT) curriculum (RCPath(UK) 2021). The new curriculum recognised that there are similar skills and knowledge required by all infection specialists and offers enhanced clinical training to trainees. However, the trainees spend less time in the diagnostic laboratories and there is less emphasis on infection prevention and control skills (Winzor and Patel 2015).

In Ireland, the Clinical Microbiologist (CM) has a key role in understanding the role of new diagnostic technologies, their characteristics (sensitivity, specificity, reliability), the problems (false-positives, false-negatives) and factors that affect those problems. They use this understanding to direct the best choice of test, both during the routine working day and during their 'out-of-hours and 'on-call' duties. The increase in antimicrobial resistance is an evolving and critical challenge in Ireland. The multi-drug resistant (MDR) organism group, Carbapenemase Producing *Enterobacterales* (CPE), was declared a National Public Health Emergency in Ireland in 2017 and is now endemic in many sites. MDR organisms, as well as globalisation and emerging pathogens, e.g. Zika virus, Coronavirus, Multidrug Resistant Tuberculosis (MDR TB), and their implications for Infection Prevention and Control, are problems which are routinely managed by the CM in Ireland, both during the routine working day and during their 'out-of-hours and 'on-call' duties.

In addition, as highlighted by the pandemic, the CM's strong understanding of the role of the patient environment, including patient equipment, room ventilation, water quality and building design, in transmission of infection, is essential to ensure both patient and staff safety. While some of this knowledge can be attained through reading and guidelines, much of this skill set is acquired through the apprentice-model of training as a member of the Clinical Microbiology and Infection Prevention and Control (IPC) team during training, gaining experience in applying these skills in a clinical context with graded responsibility. The change in the UK training were discussed with the members of the Irish Society of Clinical Microbiologists (ISCM) and members of the Specialist Training Committee (STC) for Clinical Microbiology, Faculty of Pathology, RCPI. It was agreed, therefore, that a move towards a 'Combined Infection Training' model, similar to that in the UK, was not appropriate for Clinical Microbiology trainees in Ireland.

The problem

Following on from the changes in the curriculum in Clinical Microbiology and Infection in the UK, there was a change in the structure and focus of the FRCPath exam in the UK. In Autumn 2016, the FRCPath part 1 exam was adapted to become an assessment of the Common Infection Curriculum. In 2017, the 'wet' laboratory practical component of the second part of the exam was removed. Concerns have been raised regarding the move away from the focus on laboratory skills and IPC, which are a core part of practice for the CM in Ireland.

1.1.3 Ireland - Workforce

In 2023, the Health Service Executive (HSE) National Doctors Training and Planning (NDTP) undertook an expert stakeholder informed review of Workforce Planning in the Specialities of Pathology. In the review, they looked at current unmet demand and demand growth. Two of the Pathology disciplines, Clinical Microbiology and Haematology, were identified as having the greatest unmet demand. The method undertaken in the review for Clinical Microbiology, was a simple one, using the ratio of Consultant Microbiologists to hospital bed numbers in the acute hospital services. The ratio of Consultant Microbiologists to bed numbers in Ireland was compared to that in other EU countries (Dickstein, Nir-Paz et al. 2016) to identify service gaps in CM. This work demonstrated an unmet demand of 81.6 Whole Time Equivalents (WTE) in CM, when including all hospital types (acute, elective and specialist), community and also specialist Virology services. The existing Consultant CM workforce in Ireland, reported by NDTP in the review at that time, was 58.3 WTE (HSE and NDTP 2023).

This difficulty in recruiting doctors in both Consultant and trainee positions in CM in Ireland has been long recognised by those at ground level. Many of the Consultant posts advertised fail to attract any applicants, particularly in the smaller model 2 and model 3 hospitals. There are many factors, including onerous on-call rotas, which often lead to early retirement and increased difficulty recruiting new trainees, a self-perpetuating cycle. In addition, there is a need to increase the number Consultant posts in CM, to create a workforce similar to that which is found in other EU countries.

The problem

Currently, recruitment of doctors from outside Ireland is hindered by the different training and scope of practice in CM in different countries. It is difficult to decipher objectively within the limitations of a one-hour interview when trying to appoint to these posts in Ireland.

1.1.4 Europe - the current context – influences and challenges

The changing landscape in Europe politically, economically and technologically impacts on how we deliver and assess education, currently and in the future. The FRCPath (UK) exam has long been recognised as a marker of quality assurance of training both in the UK and outside. The occurrence of Brexit in 2020 has introduced additional challenges for some countries wishing to travel to the UK for higher education and assessment (Highman, Marginson et al. 2023). For many years, trainees from around the world have travelled to the UK to sit the exam. In this context, the option of sitting an alternate, quality-assured European examination in MM could be seen as a solution.

The rapidly evolving technological developments at the intersection of education and artificial intelligence (AI), is very exciting with respect to potential delivery of education.

With AI, there is the potential to improve equitable access to high quality education (Zarei, Eftekhari Mamaghani et al. 2024). The evolution of AI is rapidly and increasingly challenging how we view assessment in education. In formative assessment, AI offers an opportunity to provide rapid timely feedback to learners and near-instant analysis of the learners performance in individual areas of the curriculum, potentially leading to targeted personalised educational activities and instruction (Krstić, Aleksić et al. 2022). In summative assessment, AI is creating more opportunities for cheating in assessment e.g. through AI-generated written essay submissions (Cotton, Cotton et al. 2023). Advances in technology also create challenges for remote invigilation of online exams (Sarrayrih and Ilyas 2013).

The war in the Ukraine, which escalated in 2022, just before the delivery of the first EEMM highlighted the challenges of maintaining education and training in conflict zones. It prompted us to address the inequalities in access to education. In April 2022, at the ESCMID council meeting, it was agreed that adjustments would be made to eligibility criteria for candidates from Ukraine to allow them to sit the EEMM. It was also agreed that the exam fee would be waived, and that sponsorship would be offered to facilitate travel to the exam centre. Medical education and assessment should be a responsive and dynamic process, reacting not only to changes in the world of infectious diseases e.g. Zika virus, SARS-CoV2, but also the greater social and political context.

1.1.5 European Scope of Practice

The specialist training, the scope of practice and the duties of the CM are somewhat different in different countries in Europe and may be quite different even between the various regions or administrative districts of one country. In some countries, the CM is

largely laboratory-based and may undertake laboratory bench-work, may oversee the laboratory management and quality assurance, and develop the laboratory standard operating procedures (SOPs). As a result, the antimicrobial stewardship, clinical liaison and infection prevention and control advice may be provided largely by phone. In other countries, the CM is more ward-based, doing regular Antimicrobial Stewardship (AMS) ward rounds, Multidisciplinary Team Meetings (MDTs) with Clinical colleagues, doing ward rounds in the Intensive Care Unit (ICU) and other specialist units e.g. orthopaedics, burns, haematology. While in most countries, the Infection Prevention and Control service is led by the CM, supported by a specialist nursing team, in other countries leading the hospital Infection Prevention and Control service may be the remit of colleagues in Infectious Diseases (ID) or Nursing.

In the 1970s, a legal mechanism was established to ensure that doctors' qualifications would be reciprocally recognised across the EU, allowing freedom of movement. In 1993, the Publications office of the European Union published the Council Directive 93/16/EEC (EU 1993) which, facilitates the free movement of doctors in Europe and the mutual recognition of their diplomas, certificates and other evidence of formal qualifications (EU 1993). In 2005, in a further effort to improve the mobility of doctors within all member states, a legal framework was created which established a mechanism of automatic mutual recognition of qualifications for medical doctors according to training requirements within all Member States; this framework is based on the length of training in the individual Specialty and the title of qualification awarded at the end of training. The publication of this Council Directive, highlighted the importance of developing a common practice of Clinical Microbiology in Europe, including a set of European Professional Standards in the speciality of CM, both for practice and for training. (Humphreys, Nagy et al. 2010)

The European Union of Medical Specialists (*UEMS: Union Européenne des Médecins Spécialistes*), is a professional organisation of Medical Specialists in the European Union. The UEMS was founded in 1958 and represents more than 50 medical disciplines and has 43 Specialist Sections, which represent independently recognised specialties. The main purposes of the UEMS are the harmonisation and improvement of the quality of medical specialists training, qualifications and practice in the European Union. In 1994, the UEMS made European level recommendations for medical training by adopting its Charter on Training of Medical Specialists (UEMS 1993). The charter which outlines the European approach to postgraduate specialist medical training requires each UEMS specialist section and European board to develop discipline specific European Standards in Medical training, which reflects current medical practice and scientific evidence.

In 2008, the UEMS established a specialist section for the discipline of MM. Prior to this, MM was a part of the UEMS section Medical Biopathology. The UEMS European Training Requirements (ETR) in the speciality of MM, European Standards of Postgraduate Medical Specialist Training, were completed in 2013 (UEMS 2013). In 2016, the MM section developed a European Curriculum which was adopted by UEMS council in 2017 (UEMS 2017). In 2015, the UEMS section of MM, at their annual general meeting, discussed the development of a European Exam in Medical Microbiology (EEMM) The EEMM would be developed as an 'exit' examination which, in conjunction with evidence of compliance with the ETR in MM and European curriculum in MM, could lead to awarding a Pan-European specialist certificate/fellowship. There is a longstanding view that assessment drives learning (Brown et al, 1997). The introduction of a EEMM was considered likely to contribute to harmonisation, reducing the differences in training and in the role and scope of practice of the Specialists in MM in Europe. This would make recruitment of doctors

from outside Ireland less difficult when trying to fill vacant posts in Ireland. The existence of a high quality EEMM would give the RCPI another option to consider as an alternative to the FRCPath (UK), as a requirement for trainees to be awarded their Certificate for Successful Completion of Specialist Training (CSCST) in Ireland.

1.2 Pilot study- a study of feasibility

1.2.1 Resources

The UEMS section of MM is a charitable organisation, with limited resources. The members are largely Consultant Microbiologists, with some representation from trainees in CM and Consultants from other UEMS sections e.g. Infectious diseases and Laboratory Medicine. Currently, 34 UEMS sections have exams. The sections with exams generally look for volunteers from their members, to engage in question-writing and review, to prepare each diet (sitting) of their exam. The exam processes and delivery vary from section to section. Some larger sections, in collaboration with their European Society, have the resources to develop their own exam platform and engage the supports of a team of administrators and educationalists. Some sections have the resources to employ an external company to manage their questions, the candidates and the exam delivery. In other sections, when the exam was developed, it was largely the work of one enthusiastic individual, without administrative or educationalist support. There is no one model that is used by all sections.

This thesis will explore the feasibility of creating a quality assured high-stakes European Examination in MM. The exploration will be undertaken by creating a pilot exam. During the creation and delivery of the pilot exam, the best approach for each step in the process will be explored, guided by the international literature on assessment quality and standards.

1.2.2 The framework

There are different frameworks which can be used to guide and evaluate an educational intervention. The models I considered were Action Research, Complex Interventions Framework and a Feasibility Study.

1.2.2.1 Action Research

Action research, which may take place in the setting of the teachers' professional environment, the classroom or school, is a research approach that may be employed to improve education. It is often a cyclical process. It may be used to address a problem that has been identified e.g. instructional practice. It has also been referred to as '*teacher research'*, because it may be undertaken by the teacher to improve their teaching by developing an "inquiry stance" to examine their own work (Cochran-Smith and Lytle 2001). Work has been done to identify quality characteristics in action research, with definitions, checklists and conceptual models for researchers to use when undertaking action research (Zuber-Skerritt and Fletcher 2007). Action research is often used as a systematic way to solve problems and create a change in education design or delivery (Yigit and Bagceci 2017).

1.2.2.2 Complex Interventions Framework

A framework was developed by the Medical Research Council (MRC), UK for developing and assessing complex interventions (Skivington, Matthews et al. 2021). The 2018 revision of the framework was extended to put more emphasis on how an intervention brings about change. The complex intervention framework divides the research into four phases; development of the intervention, feasibility, evaluation and implementation (Skivington, Matthews et al. 2021). There has been debate around whether medical education

research projects should be approached in this way, because as an intervention they involve more than one component and therefore are complex (Mattick, Barnes et al. 2012).

1.2.2.3 A Feasibility Study

The intervention in this project is a pilot exam in MM. In research, a pilot study may be considered a feasibility study (Donald 2018). Several authors have proposed checklists or tools to evaluate feasibility studies (Thabane, Ma et al. 2010, Moore, Carter et al. 2011). The purpose of a feasibility study is to evaluate the methods and the processes in the study, before embarking on the 'real study'.

1.2.2.4 The pilot exam

The purpose of this project is to develop and evaluate the processes required to deliver a pan-European Exam. Every step in the process; the curriculum model, question development, blueprinting, standard-setting and reporting the results will be informed by a review of the literature and best practice. A pilot exam will be used to simulate each step in the process and to assess if this is feasible, in our section of MM, within the current and future resources. The framework I will use to evaluate the pilot exam is based on Thabane's approach to a feasibility study (Thabane, Ma et al. 2010) and using some elements of the complex intervention framework (Skivington, Matthews et al. 2021). I chose this approach because they focus on the methods and the process required to ensure quality. I believe that the action research approach may be useful, in the future, when we undertake the next review of the European curriculum in MM.

1.3 Statement of Hypothesis & Aims

The criteria used to assess a pilot and feasibility study will be applied to the pilot European Examination in MM, to test the methods and processes and support judgements regarding the progression to the first European Examination in MM. The analyses will be used to propose criteria or a tool to evaluate a pilot exam, as a feasibility study of the assessment. Complex interventions guidance includes assessment of feasibility in its framework for developing and evaluating complex interventions. The introduction of a pan European Examination shares some similarities with the introduction of a complex intervention (Skivington, Matthews et al. 2021) and could align to the phases and core elements of complex intervention research; intervention, feasibility, evaluation and implementation (Skivington, Matthews et al. 2021). The Feasibility study may be designed to assess the intervention itself, looking at content, delivery, acceptability, likely cost effectiveness or capacity of providers to deliver the intervention, all of which are important in a high-stakes exam. The recent revision of the framework also focusses on the need to understand how interventions bring about change. In the context of the intervention, the pilot exam, one of the anticipated outcomes is progression towards harmonisation of training and practice of Clinical Microbiology in Europe. Furthermore, the demonstration of feasibility and a positive evaluation may lead to development of an assessment that could be considered as an alternative option for Irish trainees in HST in Clinical Microbiology.

The main aims / stages of this feasibility study were as follows:

1. Development & standard psychometric evaluation of a Pilot European Assessment in Clinical Microbiology (Chapters 2 & 3).
2. Development of further tools to evaluate the assessment (Chapter 4)
    a. Development of an assessment tool to evaluate the trainee experience of the European Exam in Clinical Microbiology.
    b. To assess the pilot examination against the criteria used in a pilot and feasibility study, modifying these criteria as appropriate, to create a tool to assess a pilot examination.
3. Evaluation of the exam feasibility and assessment (considering points 1 & 2 above) to see if modifications may be required for implementation of a pan-European Assessment in Clinical Microbiology with respect to training and scope of practice in Europe.

Chapter 2: Creating a quality assured high-stakes exam

2.1 Background

2.1.1 Scope of practice

As previously described, the role and scope of practice of the CM varies from country to country. In some countries, the CM is largely laboratory based, while in other countries, the CM is more ward-based and patient facing (Humphreys, Nagy et al. 2010).

The role of the CM has been evolving over the past five decades. Historically, in the UK, the CM directed a lot of the laboratory bench work, however, since the 1970s, the technical and scientific staff in the laboratory have largely taken on this role and the CMs focus is on reporting and interpreting the results and liaising with clinical colleagues, the "*interface between the laboratory and the clinicians*" (Scott 2000). While we know that a similar change has happened, or is happening, in other countries, it is difficult to know to what extent the scope of practice has adjusted.

A study undertaken almost 30 years ago (Mehtar 1995), demonstrated that there are at least three main areas of practice for the CM; management of infection (which includes clinical interpretation of results), Laboratory management and quality and Infection Prevention and Control. The local situation in which the CM is practising determines when any one of these areas may take priority over the other. The development and access to new technologies such as Polymerase Chain Reaction (PCR), automation and sequence-based identification of micro-organisms is also influencing the role and practice of the CM (Didier Raoult 2004). This is also the case in the United States (US), where the role of the

CM Director has undergone significant changes with the introduction of new technology (Thomson and Doern 2011).

The role of the Microbiologist can even vary within one country, depending on the size and type of hospital they practice in. Some CM may provide a service to more than one hospital, which will also impact on their role and daily practice. It is difficult to provide a ward-based clinical liaison service if you are not based on the same site as the patient (Mehtar 1995). Leadership of the Infection Control service in the hospital is the role of the CM in many countries. In some countries, the CM may also have a role in Infection Prevention and Control in the community (Scott 2000).

A survey of CMs practice, undertaken in South West England, showed considerable variation in time spent carrying out specific tasks e.g. time spent on Intensive Care Unit (ICU) ward rounds varied from 1-10 hours per week, Infection Prevention and Control (IPC) activities ranged from 2-16 hours per week, in some sites more IPC duties were undertaken by the IPC nurse and there were significant differences in the time devoted to public health microbiology (Riordan, Cartwright et al. 2002). The role and scope of practice of the CM varies and is dependent on many factors including the demand created by clinical colleagues, the support for the service provided by other healthcare professionals (Medical, Nursing, Scientific), the type of hospital, and the country of practice. The creation of the European Core Curriculum in MM (UEMS 2017) will direct training to develop skills in the three core areas of practice.

> "*Assessment defines what students regard as important, how they spend their time and how they come to see themselves as students and then as graduates*".
>
> (Brown, Bull et al. 1997)

The development of a European assessment in Clinical Microbiology, based on an agreed European core curriculum, should encourage training and study in all areas of the curriculum and lead to better harmonisation of scope of practice.

2.1.2 Curriculum design

"*Assessment is the tail that wags the curriculum dog*".

This oft-quoted phrase from Eleanore Hargreaves (Hargreaves 1989), recognises the importance of assessment; that it can drive what the learner will learn and the teacher will teach. With this understanding, the importance of quality and rigor in the assessment process, cannot be over-emphasised, particularly in high-stakes assessment. But first, we must consider the '*dog*' - the curriculum.

Quality assurance is a vital part of assessment, particularly for high-stakes examinations. A high- stakes examination is one which is '*strongly consequential for student progression*' (French, Dickerson et al. 2024). The EEMM is designed as an 'exit' exam, which the trainee may sit when they have completed much of their specialist training. High-stakes, exit exams in medicine are seen as a means to reassure stakeholders (e.g. trainees, trainers and the wider society) that the healthcare provider is competent at the end of a period of training (Alsulimani 2021).

Development of any assessment needs to start with the curriculum. The days when the teacher created the curriculum in the manner of a magician pulling a rabbit from a hat are gone, a curriculum should be carefully planned and focussed on what will happen in the teaching programme (Harden 1986).

The first consideration with respect to the curriculum is to identify the 'need' (Harden 1986). The need may be determined by the teachers, the learners, the employers and in

medical education, we should also consider the societal need, the social responsibility to train a specialist that will meet the needs of the population it serves (Harden 1986). The curriculum needs may be identified by the stakeholders e.g. the public, patients and employers. They may be identified by the '*wise man*' approach, where senior practitioners discuss and agree (Dunn, Hamilton et al. 1985). Other approaches include looking at '*errors in practice*', looking at '*critical incidents'*, undertaking '*task analysis'* and assessing the role of the specialist (Dunn, Hamilton et al. 1985). In reality, assessing curriculum needs is an iterative process, informed and reinformed by a combination of these approaches.

Curriculum theory is a longstanding area of study within education, with many varying views on the definition of a curriculum (Quinn 2000). Curriculum may be seen as 'opportunities' for students, created by their engagement with education, or may be seen as objectives i.e. what behavioural changes will occur in the student. Curriculum may be seen simply as subject matter, made available to the student, or as the 'student experience', of all the learning inside and outside the education facility, including social opportunities (Tyler 1949, Kerr 1968, Bell 1973, Quinn 2000). The curriculum can also be viewed in a more generic way as an overall plan for learning opportunities (Saylor, Alexander et al. 1981), or the 'action research' approach, focused on teaching and learning goals and implementation of the curriculum in practice (Stenhouse 1975).

When a curriculum is developed, there are many options for how the development is approached, but Harden describes the four most common (Harden 1986). The *architect approach* is all about the plans with a focus on the learners outcomes, the *mechanic approach* focusses on the workings of the curriculum, teaching method and strategy, the

cookbook approach focus is the ingredients or content, and finally, 'the *railway timetable approach*' considers the timetable of teaching events (Harden 1986).

The approach to developing the curriculum will be influenced by the curriculum model that is used to plan the curriculum, by providing the curriculum developers with a system that can be transparently used to explain their decisions with respect to their approach to teaching activities and assessment (Ornstein and Hunkins 2009). There are many papers describing approaches to curriculum design, with two commonly described, polarised, curriculum models; the Product Model and the Process Model (Quinn 2000, Neary 2003). Curriculum models may also be described as technical or non-technical; these terms are similar to the product/process divisional, with the technical curriculum focusing on curriculum as a blueprint to structure the learning environment and the non-technical curriculum focusing on the learner (Munna and Kalam 2021).

The Process Model of curriculum

In the process curriculum model, the focus is on the journey, rather than the destination. There is less focus on the final grades that the learner will achieve and more focus on the educational experience and the "value added" in teaching (Tummons 2012). The process model is linked to cognitivist pedagogical thought (Piaget 1952), which considers the intellectual development of the learner and focusses on developing the learners' cognitive skills. The focus is on the internal cognitive processes, shifting focus away from visible actions. Stenhouse challenged the traditional view of a curriculum as fixed content, emphasising the importance of flexibility in education and responding to the students' needs and interests (Stenhouse 1975).

In the process model, the curriculum content has its own value and therefore is not selected based on learning objectives. Learning activities have their own value (Gatawa 1990). The teacher is seen as a facilitator, supporting the students' learning (Lynch and Smith 2011). The learners are considered active creators of knowledge (Barnett, Parry et al. 2001) and the curriculum is viewed as a dynamic process (Schön 1987). Knight supports the process model, which he views as starting from complexity theory and complexity in learning, in which the ingredients are what are important i.e. the processes, messages and conditions (Knight 2001). He believes that the product model is more suited to a training culture than an educational one (Knight 2001).

William Spady, among others, criticised what he called the 'calendar-based model', the idea that if you simply teach students using a set amount of material, for a set time period, that they will *de facto* learn what they need to know (Spady 1988).

The Product (Behavioural Objectives) Model of curriculum.

The Product Model is an output model which emphasises the achievement of specific objectives by the learner. This curriculum model has really existed for over 500 years. Any apprenticeship training model where the goal is to attain a practical skill is an outcome-based system e.g. training in the craft guilds of the Middle Ages (Spady 1993).

Many believe it may have its foundations in the behaviourist theory of instruction, starting with the early work of Ivan Pavlov, the Russian Physiologist, best known for his work conditioning dogs (Pavlov 1927). The behaviourist John B. Watson conducted similar, and somewhat controversial, studies, with a nine-month-old boy, exposing him to stimuli and observing the boy's reaction. He demonstrated that specific learning experiences can alter behaviour, in his work with 'little Albert' (Watson and Rayner 1920). Burrhus Frederic Skinner introduced the concept of 'operant conditioning'; that a learner's behaviour is

influenced by the response to the behaviour, the response being positive or negative reinforcement of the behaviour. He emphasised the importance of both observable behaviour and the internal mental thoughts and processes (Skinner 1948). Behaviourism recognised the importance of clear identification and measurement of learning, and the need to produce observable and quantifiable outcomes.

While the origins may lie in the work of the behaviourists, most would agree, however, that the foundations and core principles of the product model of curriculum lie with work by Ralph Tyler (Tyler 1949) and later, Benjamin Bloom (Bloom and Krathwohl 1956). Tyler emphasised the importance that any objectives stated should centre around a change in the behaviour of the student, and that the behaviours should be both observable and measurable. He believed that when developing a curriculum, four questions should be answered; what is the educational purpose (objective) of the curriculum, what learning activities (content) will achieve the objectives, how can the learning activities be organised (method), and how can the teacher evaluate whether the learning experience has been effective (evaluation) (Tyler 1949). This model of curriculum, consisting of four components (objectives, content, methods and evaluation) has influenced education throughout the world.

Outcome-Based Education (OBE)

In the early 1990s, William Spady proposed Outcome-Based Education (OBE), a pedagogical model in which the curriculum is restructured to describe what the student is expected to know, or be able to do (Spady 1994). The learning that the student attains (the product) is more important than how it is attained (the process). The learning the student attains is expressed as a 'Learning Outcome'. The idea of writing learning outcomes can be traced back to Robert Mager who suggested writing specific statements about outcomes that are observable (Mager 1975). There are many definitions of a learning outcome, the definition from the European Credit Transfer and Accumulation System (ECTS) is:

> "*Learning outcomes (are) statements of what a learner is expected to know, understand and/or be able to demonstrate after a completion of a process of learning.*" (ECTS 2005).

OBE promotes a self-directed approach to learning, an important skill for 'life-long' learning, giving the learner the necessary skills for their continuous professional development. In the OBE curriculum, the focus is on the learner, which in the context of medical training, the type of doctor that learner will be (Harden, Crosby et al. 1999). The OBE model of curriculum design, creates greater curriculum transparency, allowing the content of a curriculum from one programme, institution or country to be easily compared to that of another (Lennon 2018).

European Higher Education Area (EHEA)

In 1998, the University of Paris celebrated its 800-year anniversary. Four education ministers attending the event, representing France, Germany, Italy and the UK, made a joint declaration to facilitate the harmonisation of the European Higher Education system, "*The Sorbonne Declara*tion" (EHEA 1998). The aim being to create a common frame of reference within the intended European Higher Education Area (EHEA) which would make it easier for students, graduates and staff to train and work between different countries.

The following year, in June 1999, the Ministers of Education of twenty-nine countries came together in a voluntary process to create the European Higher Education Area (EHEA). They formally launched the Bologna Process with the Bologna Declaration of 1999, which is now one of the main voluntary processes at European level, defining the European Higher Education Area (EHEA) (EHEA 1999). Today, forty-eight countries, including Ireland, are part of this process and are adopting reforms on higher education.

The Bologna process continues to propose a series of reforms to make European Higher Education more comparable, their aim being to facilitate easier mobility of staff and students between faculties, and to support cross-border academic co-operation, with the mutual recognition of qualifications earned abroad and time spent studying abroad. Quality assurance and enhancing the quality of learning and teaching is a core mission of the Bologna process. The Bologna process describes a number of 'action lines', in which an important role should be played by learning outcomes (Adam 2004). All programmes in third level institutions throughout the EHEA should have curricula which are designed based on learning outcomes, not on the lecture content of professors, by 2010.

Learning outcomes form a critical part of the Bologna education reform. In the European Higher Education Area (EHEA), learning outcomes were first mentioned in the 2003 Berlin

Communiqué, not as a stand-alone goal, but as one tool to achieve comparable degrees. (Communiqué 2003). In 2005, *The Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG)* (ESG 2015) were adopted. Since then, much progress has been made including the promotion of the use of learning outcomes, supporting the shift towards student-centred learning and teaching (Communiqué 2005). Internationally, learning outcomes can promote the mobility of both students and those seeking employment by facilitating credit transfer and the recognition of qualifications. The creation of a system of easily comparable degrees by establishing a system of credits, the European Credit Transfer and Accumulation System (ECTS) is one of the most important tools recognised by the Bologna Process (Commission, Directorate-General for Education et al. 2015).

In the context of the Bologna process and reforms in European Higher education, coupled with the evidence that in OBE, the learning outcomes provide a framework to which the teaching, learning and assessment can be aligned and therefore improving quality assurance, the European curriculum in MM was converted to an outcome-based curriculum to facilitate the development of the examination.

The flexibility which the outcome-based curriculum affords, is in specifying what the learner is expected to know, understand and / or be able to demonstrate after completion of a process of learning, but not the form of course delivery or the educational strategy (Harden, Crosby et al. 1999). It is believed that OBE links to mastery learning (McNeir 1993), as compared to criticisms of the 'calendar-based model', the idea that if you teach students using a set amount of material for a set time period, that they will learn what they need to know (Spady 1988). Mastery learning is the standard expected of the specialist in Clinical Microbiology.

The UEMS section MM curriculum, which was ratified by council in 2017, had a more cookbook approach, focussed on the 'ingredients' (Harden 1986) and was a more process or non-technical model, where the curriculum content has its own value (Knight 2001). Recent years have seen a move away from the non-technical or process model, and towards the technical, Education Product model, often described as Outcome-Based Education (OBE), which has been described as leading to mastery learning (McNeir 1993).

Writing Learning Outcomes:

In 1948, at a physiological convention in Boston, Benjamin Bloom led a group of educational psychologists in developing a classification system, intended as a framework for communicating education goals; the taxonomy of educational objectives. The taxonomy divided educational objectives into three original domains; cognitive, psychomotor and affective (Bloom and Krathwohl 1956). The first, the cognitive domain, focussed on knowledge and intellect e.g. understanding, analysing, evaluating. The second was the psychomotor domain, which focussed on acquisition of a skill e.g. performing, assembling, dismantling. The third domain was the affective domain, which considers attitudes, feelings and values e.g. appreciating, accepting (Bloom and Krathwohl 1956). Bloom recognised that there should be an ascending order of complexity within each domain, creating a framework whereby the learner can build on prior learning, moving to the higher or more complex level of learning. In Bloom's Taxonomy of Education Objectives, six levels of cognitive development are described; (Bloom and Krathwohl 1956).

Level 1 Knowledge (knows)

Level 2 Comprehension (Understands)

Level 3 Application (Applies knowledge)

Level 4 Analysis (analyses knowledge)

Level 5 Synthesis (Uses knowledge to reach a conclusion, account for)

Level 6 Evaluation (Uses knowledge to appraise, judge)

The taxonomy was revised in 2001 by Anderson and Krathwohl, to include at the highest level, creativity (Anderson, Krathwohl et al. 2001), however the original work of Bloom and his colleagues still remains the most widely referenced in the educational literature.

The second category for grouping instructional objectives is the psychomotor domain or 'Skill' domain, which focuses on physical skills involving co-ordination between the brain and muscle activity. Developments in this domain, in the educational field, seem to be less, and while Bloom and colleagues did include this domain in their 1956 publication of taxonomy, they did not publish specific levels in this domain because felt that they lacked the experience of teaching psychomotor skills. There are however other educators who have developed a taxonomy to describe the skills domain. One of the best known was developed by Harrow, which focuses on agility, physical fitness and dexterity. It was developed for children with special needs and is often used in the context of physical education (Harrow 1972). Dave developed a taxonomy which focuses on the ability to co-ordinate motor skills. The descriptions of the levels, below, are simple and easy to apply in the medical education environment. (Dave 1970)

Level 1 Imitation (Observing and copying someone else).

Level 2 Manipulation (Guided via instruction to perform a skill).

Level 3 Precision (Accuracy, proportion and exactness exist in the skill performance without the presence of the original source).

Level 4 Articulation (Two or more skills combined, sequenced, and performed consistently).

Level 5 Naturalisation (Two or more skills combined, sequenced, and performed consistently and with ease. The performance is automatic with little physical or mental exertion).

The third category of affective domain encompasses attitudes, appreciations, values, and emotions – although highly important in education, this is one of the hardest to assess the student. In 1964, eight years after the publication of Bloom's Taxonomy, a scale was produced to describe five levels in the affective domain (Krathwohl, Bloom et al. 1964). Krathwohl's affective domain taxonomy is perhaps the best known of any of the affective taxonomies. The taxonomy is ordered according to the principle of internalisation. Internalisation refers to the process whereby a person's affect toward an object passes from a general awareness level to a point where the affect is 'internalised' and consistently guides or controls the person's behaviour. The levels of affective domain form a continuum from simple awareness and acceptance, then as attitudes become part of an individual's practicing value system, to the highest level, internalisation.

Level 1 Receiving (to differentiate, to accept, to listen (for), to respond to).

Level 2 Responding (to comply with, to follow, to commend, to volunteer, to acclaim.)

Level 3 Valuing (valuing certain ideas, to support, to debate.)

Level 4 Organisation (relating the new value to that one already holds, to discuss, to theorize, to formulate, to balance, to examine.)

Level 5 Characterisation (acting consistently in accordance with the values the individual has internalised, to revise, to manage, to resolve.

Professor John Biggs introduced us to the concept of 'constructive alignment, which has been one of the key drivers in reforming curriculum in education, demonstrated by its influence on Quality and Qualification systems all over the world (Biggs 1999). Successful constructive alignment requires that three tasks occur; clearly specified learning outcomes, teaching and learning activities to achieve the outcomes and the assessment to demonstrate that they have been achieved (Biggs 2003). For quality assurance and planning of assessment, specifying the learning outcomes, also means that the standard that the learner should achieve can be stated, rather than the time the learner takes to achieve it (Harden, Davis et al. 1997). This is critical for a high-stakes, criterion-referenced assessment. When designing an assessment, learning outcomes clearly show how well the assessment covers all the relevant parts of the curriculum, the content validity which enhances credibility and reduces bias (Hopkins 1998). To facilitate the development of assessment and move the curriculum closer on the continuum to the product model, it was agreed that the existing European Curriculum in MM would adapt to an OBE curriculum model.

2.1.3 Assessment design

The purpose of assessment is to use a specific sample to make a generalisation around the level and extent of the learners learning and skills (Norcini, Anderson et al. 2011). In order to do this, assessments need to be reliable, feasible, fair and should also benefit learning (Van Der Vleuten 1996).

Miller's pyramid (Figure 2.1) is a useful tool for planning and guiding assessment of a curriculum, providing hierarchical framework for assessing learners level of knowledge and skills (Miller 1990). Starting at the bottom, is the most basic level of knowledge, 'knows', the second level is where the learner demonstrates some competence in application of knowledge, 'knows how'. The third level assesses the learner's performance in demonstrating their skills, 'shows how', and the final level of the framework is 'does', where the learner now performs their role or task competently.

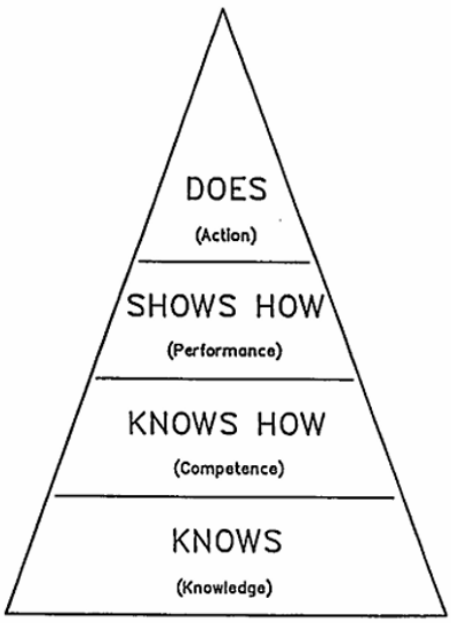


Figure 1. **Framework for clinical assessment.**

Figure 2.1 Millers Pyramid

Millers Pyramid provides hierarchical framework for assessing learners level of knowledge and skills (Miller 1990). At the bottom, is the most basic level of knowledge, 'knows', the second level is application of knowledge, 'knows how'. The third level assesses, 'shows how', and the final level of the framework is 'does', where the learner now performs their role or task competently.

Figure 2.1 Miller's Pyramid

Miller starts by acknowledging that

> "*…no single assessment method can provide all the data required for judgement of anything so complex as the delivery of professional services by a successful physician*" (Miller 1990)

and subsequent authors have suggested various assessment modalities appropriate to each level of the pyramid. Alternatively, Bloom's Taxonomy can be used to plan the level of performance or various domains to be evaluated in an assessment (Dzara and Gooding 2022)

Formative assessment

Formative assessment is an important part of assessment in OBE, and works best if it is embedded within the training process and involves specific, timely feedback (Norcini, Anderson et al. 2011). As part of the application process to sit the European exam in CM, trainees will be required to submit references confirming that they are in a training post, under the supervision of a training director or equivalent. It is expected that formative assessment will be undertaken by the trainee's local supervisor while on the training programme. The draft framework for the EEMM, recommends that trainees use a logbook or checklist to document and reflect on learning experiences (Snadden and Thomas 1998). The logbook or checklist is also very useful if the trainee will train over multiple training sites. (Schüttpelz-Brauns, Narciss et al. 2016). A survey looking at training in Clinical Microbiology in Europe showed that in all trainees from the countries that responded to the survey recorded their training activities in a logbook or training record (Doyle, Boyle et al. 2021). A variety of tools have been created to facilitate formative assessment e.g. mini-clinical evaluative exercise (mini-CEX) and Case-based discussion (CbD). The trainer

provides feedback either during or after the assessment (van der Vleuten and Verhoeven 2013). Formative assessment is an opportunity to assess the higher levels of Miller's pyramid e.g. Case-based discussion may assess level 2, 'knows how' and mini–Clinical Examination or Directly Observed Procedures may be used to assess level 3 or 4, 'shows how' and 'does'.

The survey looking at training in Clinical Microbiology in Europe, showed that in over 60% of countries surveyed, trainees undergo formative work-place-based assessments at their training site (Doyle, Boyle et al. 2021).

Quality in Assessment

Several criteria have been described that are important when designing an assessment to ensure good, quality assessment e.g. validity, reliability, discrimination and feasibility (Norcini, Anderson et al. 2011, Norcini, Anderson et al. 2018).

"*A test is valid if it measures what it purports to measure*" (Kelley 1927).

Test validity is essential for quality assurance in assessment. The concept of validity in a test is not recent but was formulated by Kelley in 1927. He identified two common fallacies in assessment; one was that two tests that you claim are testing the same thing are actually testing the same thing, and the other that two tests that you claim are testing different things are actually testing different things. Therefore, validation was establishing that the test is assessing what you claim it is assessing (Kelley 1927). It is, however, impossible to prove that a test is valid, or prove that it is measuring what we claim that it is measuring (Messick 1989), but instead validity is about the interpretation of the test, and the inferences we make based on the test scores (Wiliam 2008).

Constructive alignment is the matching or aligning of the assessment to the learning outcomes in the curriculum, taking into consideration the *level* of attainment you expect the learner to achieve for the learning outcome The level can be specified using a taxonomy such as Bloom's Taxonomy (Bloom and Krathwohl 1956) and the mapping process itself is referred to as blueprinting. Validity of assessment is closely associated with 'constructive alignment' (Quinn 1997).

In Validity theory, validity is the property of the test score *interpretation or meaning*, not about the property of the test (Borsboom, Mellenbergh et al. 2004). Historically, three types of validity are described, content validity, construct validity and criterion validity (which includes concurrent and predictive), however Samuel Messick believes this view is incomplete, because it does not consider the implications of the score in terms of how it is used, or the social consequences of using the score (Messick 1995). Messick describes a new *'unified'* concept of validity. In validity as a unified concept, there are six aspects to construct validity: content, substantive, structural, generalisability, external and consequential. These function as standards for assessment (Messick 1989) (Messick 1995).

Clearly, there are many different types of validity to consider when designing an assessment and choosing the assessment methods (Gronlund and Linn 1990). *Construct validity* is when the assessment tool assesses what it was designed to assess, e.g. assess at the correct level of Miller's pyramid. *Construct underrepresentation* occurs if the sampling is too narrow and *construct-irrelevance variance* if the assessment is too broad (Messick 1995). *Content validity* is achieved by ensuring that the sampling in the assessment is adequate, that a sufficient number of learning outcomes are assessed from throughout the curriculum. Blueprinting demonstrates how this can be achieved (Messick 1989). *Predictive validity* is the ability of the test to predict how the student will continue

to perform in the future. *Concurrent validity* is that the test and criterion data are gathered at similar points in time and *substantive validity* is the observed inconsistencies in test responses (Messick 1995). *Structural validity* is the extent to which the scores of the measurement instrument are an adequate reflection of the dimensionality of the construct being measured (De Vet, Terwee et al. 2011). *Generalisability* is the extent to which the scores and their interpretations can be generalised across different population groups and settings (Cook and Campbell 1979). *External validity* is convergent evidence from comparison with another method (Messick 1995) and *consequential validity* is the implications of acting on the score interpretations (Messick 1995).

The holistic approach to validity is as "*an overall evaluative judgement of the adequacy and appropriateness of inferences and actions based on test scores*" (Messick 1988) (Wiliam 2008). Robert Ebel took a different view, wherein that because educational assessments were not psychological tests, validity has to be built into the test with good quality item-writing and blueprinting (Ebel 1983).

Reliability

Test reliability is also important for quality assurance in the assessment (Gronlund and Linn 1990). Reliability in a test ensures that the test consistently measures or tests what it was designed to assess. A reliable test should give the same assessment of the learner's ability even if the test was marked by a different rater (examiner). Unreliability can occur for a number of reasons including errors in marking, different examiners with different ideas of what the learner standard should be, poor design of the assessment technique, and too many choices on the paper (Downing 2004). Reliability is often achieved at the 'expense' of validity, so when we design an assessment, in addition to considering the feasibility, we

are often trying to strike the balance between the two (Wass, Van der Vleuten et al. 2001). Reliability is discussed in further depth in Chapter 3 of the thesis.

Another consideration when designing a test is to consider the *discrimination* of the assessment and of each item in the assessment. Are there enough 'difficult' items and 'easy' items in the assessment to achieve a spread of test scores, demonstrating the candidates' range of ability (Schuwirth and van der Vleuten 2020). This will also be discussed further in Chapter 3.

When designing an assessment, particularly for candidates from different countries, the test items should be scrutinised for content that might introduce bias e.g. Language or context. It is useful to have representation from different countries to minimise this type of bias (Roberts, Sarangi et al. 2000), and is an important consideration when designing the European assessment. The examiners should also be mindful of inclusivity and equity when designing test items, which is particularly relevant for exams such as OSCEs and OSPEs (Ibrahim, Juve et al. 2023).

Finally, the other important consideration is whether the test is practical or *feasible.* This includes factors such as the cost of the test, the time taken to conduct the test and the examiners and resources required to deliver the test (Education 2010).

Assessment method

The next step in designing an examination is to consider the test method, while remaining cognisant of the need for validity, reliability, discrimination and feasibility.

The type of assessment method used can also affect how students learn, whether they strive simply to recall facts, 'surface learning', or whether they develop a deeper understanding, 'deep learning', through application of knowledge (Dinsmore and

Alexander 2012). The deeper level of learning is what is considered successful learning in the Bologna declaration (Asikainen, Parpala et al. 2014).

Traditionally, many high-stakes exams have included an essay and/or short answer questions. The examinations were often arbitrary and the standard chosen to select the passing candidates were often subjective (Norcini, Anderson et al. 2011). Essays can be very useful at assessing deeper knowledge and the higher levels of Miller's pyramid, but due to the limited sampling and inter-rater variability, their validity and reliability can be poor (Wass, Van der Vleuten et al. 2001). Short answer questions have better validity as they will sample the curriculum more widely, however, the inter-rater reliability can be poor. In the context of the European exam, the marking for essays and short answer questions is very time-consuming, so it would not be feasible. Furthermore, for many candidates, English will not be their first language and this might disadvantage some students as marks may be affected by the students' grammatically correct and expressive prose (Roberts, Sarangi et al. 2000).

Modified Essay Question (MEQ) and Patient Management Problems (PMP), while better assessments of application of knowledge, are hampered by the difficulty with getting consensus from all the judges. An additional factor, particularly in PMPs, is the risk of 'cueing' (prompting the learner) when the test is delivered on a written paper. These factors may affect the reliability of the assessment, but can be somewhat overcome with training and carefully designed, agreed, clear scoring keys (Feletti 1980).

Objective tests were initially developed to eliminate the inter-rater reliability seen in the essay test (Norcini, Anderson et al. 2011). Commonly used written objective tests are those in which the student is given the answer in a list of options. e.g. Multiple-choice questions (MCQ), such as Single Best Answer (SBA) or Extended matching Questions

(EMQ). True/False is another option, but the student may guess the answer, so some assessment employ a negative-marking strategy for incorrect answers to deter guessing (Case and Swanson 1993, Jennings 2012). They usually have very high content validity because they can include many items and therefore sample widely throughout the curriculum. SBA may be vulnerable to *cueing*, where a candidate recognises the current option but could not have answered it if they had not seen it (Schuwirth, van der Vleuten et al. 1996). In EMQ, in which several questions are asked with the same long list of possible answers, cueing is minimised (Schuwirth, van der Vleuten et al. 1996). They have good feasibility, because marking may be done by means of optical scanning sheets and a computer (Chai 2016). The other advantage of objective tests is that detailed psychometric evaluation can be performed on the test items, providing validity evidence, as will be further discussed in Chapter 3 (De Champlain 2010). The ability of these test items to discriminate between high-performing and lower-performing students is strongly dependant on the quality of the questions (Schuwirth and van der Vleuten 2004). They can evaluate knowledge, at the lowest level of Miller's pyramid, and when well-constructed, may also evaluate the application of knowledge (Case and Swanson 2002). However, the question-writing is very time consuming, and it can be challenging to write questions that will test 'deeper' learning.

Testing the higher levels of Miller's pyramid is more challenging as it requires direct observation of candidates by experienced judges. In the past, oral examinations were popular, but while they have good validity, they tend to have poor reliability and consistency, are difficult to standardise, and are time-consuming (Wass, Van der Vleuten et al. 2001). Significant resources are required to improve their reliability (Wass, Wakeford et al. 2003). The use of an Objective Structured Clinical Examination (OSCE) (Harden and

Gleeson 1979) or an Objective Structured Practical Examination (OSPE) (Belovarac, Zabar et al. 2021) facilitates assessment of areas of the curriculum in the psychomotor and affective domains. However, candidates may not perform consistently or to the same standard from task to task, and broad sampling across stations is essential to assess overall competence reliably (Wass, Van der Vleuten et al. 2001). Another consideration is the reliability and inter-rater variability. While this is less important for instruction or formative assessments, this has implications for the defensibility of high-stakes, licensing summative assessments (Van der Vleuten and Swanson 1990, Swanson and van der Vleuten 2013).

Workplace-based assessments are an opportunity to observe the trainee and give immediate feedback following formative assessment (Norcini and Burch 2007). Many formative assessments have been devised, including the Directly Observed Procedure (DOPs) assessment, where a learner is observed and assessed undertaking specific tasks. The use of a simulated patient is also an option. The challenge in this type of assessment is appropriate sampling, as one single observed encounter with a real or simulated patient cannot be relied upon to assess with confidence the learner's ability across cases or domains. This is known as case-specificity, and is a threat to both validity and reliability (Downing 2004, Schuwirth and van der Vleuten 2004). The evidence to support this type of work-place based assessment is inconsistent, but this probably relates more to feasibility, and the time available to perform the assessments (Norcini, Anderson et al. 2011).

Computer-based testing and computer-based simulations have increased in popularity since the global pandemic and when well-designed, may offer a robust test of the learners' management of complex clinical problems (Guagnano, Merlitti et al. 2002). This

assessment format may provide opportunities for assessment of the higher levels of Miller's pyramid in the future (Miller 1990) (Ward, Muckle et al. 2019).

Assessment Blueprinting

Assessment requires a 'well thought-out plan' to ensure that it is designed to confirm acquisition of the curriculum learning outcomes, and so the curriculum, teaching methods and programme of assessment should all be aligned (Biggs 2003, Raymond and Grande 2019). The specification of the learning outcomes is important for effective planning and implementation of the assessment (Harden, Davis et al. 1997). Blueprinting of assessments against the curriculum is essential in a systematic approach to the design of a programme of assessment, described as *mapping* (Hamdy 2006). It is a sampling strategy that avoids undue duplication and gaps, ensuring a good spread of content and skill coverage, to deliver a valid and reliable test (Biggs 2003).

In addition to a blueprint for a programme of assessment, each test should have a blueprint showing how curricular content will be covered and sampled in any individual assessment. Blueprinting can be divided into different levels; Macro-level blueprinting provides an overview of all assessment techniques, formative and summative, that will be used to assess the curriculum (programme) learning outcomes. Meso-level blueprinting can be used to blueprint the examination topics. Micro-level blueprinting considers the test item and the learning outcome(s) it assesses. Mark Raymond describes four stages in developing an effective test blueprint; firstly, identify the major knowledge and skills domains, then identify the learning outcomes to be assessed within each domain, determine the assessment format and specify the weight that will be given to each content category (Raymond and Grande 2019).

There is no ‘best way’ to construct the blueprint, and method used must ensure that the test content aligns with the curriculum (Notar, Zuelke et al. 2004). When creating a test blueprint, the test blueprint may be content-oriented, that is describing the test in terms of the specific content of the course or module that will be tested (Raymond and Grande 2019). A test blue-print may also be ‘process-oriented’, for example if it is assessing procedural skills, and can also be very useful for clinical skills where you may wish to assess in more than one domain (Raymond and Grande 2019). These two frameworks can be merged into a ‘content-by-process’ matrix, which offers great flexibility and is often used in the increasingly popular “integrated curriculum” (Brauer and Ferguson 2015).

Validity is required for every assessment and implies that if the candidate achieves a passing standard, that they have achieved the minimum competence required, as described in the learning outcomes (Coderre, Woloschuk et al. 2009). As discussed earlier, there are two major threats to validity of an assessment, which can be reduced by blueprinting, *construct under representation* and *construct-irrelevance variance*. (Messick 1989).

Blueprinting will result in a pre-established table of content and processes which is weighted to reflect the breadth and relative importance. Content areas of the curriculum can be weighted, so priority can be given to more ‘important’ areas of the curriculum. (Coderre, Woloschuk et al. 2009). Defining which elements of content that should be prioritised can be difficult, but it is also important that to define the purpose and focus of the assessment. When creating the assessment blueprint for the European Exam, consideration was given to the future job of the doctor, using the scope of practice survey to ensure that the topic areas assessed are reflective of the typical practice of a CM, but also respecting the content of the European Curriculum.

Assessment - Curriculum - Learning

Assessment, curriculum and learning are undoubtably inter-linked. Newble describes how a curriculum reform which was not matched by an assessment reform affected student behaviour. Ward-based teaching was introduced to replace didactic lectures. The assessment methods did not support the curricular reform. They were more suited to didactic teaching, so the students stopped going to the wards and instead requested more didactic teaching (Newble 1998). In this case, there was an additional unintended effect, undermining the institutional mission, by using an assessment method that was not aligned (Trigwell 2001). This highlights the close and inter-dependant relationship between curriculum and assessment; assessment is not just a measurement, it is the '*tail that wags the dog'*.

2.2 Aims

The UEMS section MM Curriculum was ratified by Council in 2017. The curriculum design was a blend of a content model and product (objectives) model (UEMS 2017). As the first step in creating the EEMM, the aims are as follows:

1. Assess scope of practice in Europe, through a survey.
2. Modify the 2017 MM curriculum to describe the curriculum learning outcomes
3. Create a blueprint for the exam
4. Create the test for a pilot exam

2.3 Methods

2.3.1 Scope of practice survey

Ethics approval was granted by the Southeast Regional Ethics Committee, Ireland

The survey was created on SurveyMonkey and had sixty-nine questions (Appendix 2). Questions concerning training and current scope of practice were entered into SurveyMonkey, and a link sent members of the UEMS section MM board. The survey used a Likert scale to determine if specific tasks were the remit of the MM; 'always', >75%, 50-75%, 25-50% and 1-25% of the time or 'never'.

The survey was circulated to members of the UEMS section MM in 2019 and again in 2022, with reminders when necessary. Participants were sent a Participant Information Leaflet, which they acknowledged receipt of before commencing the survey. Participants completed a consent form prior to completing the survey. The questionnaire was anonymous, but each participant was asked to identify the country in which they were

working. The UEMS section MM board currently has 51 members with representation from 27 countries in Europe. Four members are liaison officers representing UEMS section of laboratory medicine, UEMS section of Infectious diseases and ESCMID. There is one trainee representative. All other members are Consultant CM.

2.3.2 Curriculum and Blueprint

The European curriculum in MM was reviewed (UEMS 2017). The curriculum is divided into nine sections. To facilitate the development of assessment, each section was reviewed to see how the content could be adapted to describe the learning outcomes for each section.

I reviewed the literature on writing learning outcomes, specifically looking at taxonomy for learning outcomes in the domains of knowledge, skills and attitudes. When writing the learning outcomes, I primarily referred to resources from two universities in Ireland, the University of Limerick and University College Cork (Kennedy, Hyland et al. 2007, O' Brien 2008). The approach described by Kennedy *et al* in their paper, 'Implementing Bologna in your Institution' (Kennedy, Hyland et al. 2007), starts with the work of Bloom (Bloom and Krathwohl 1956) and combines Fry's recommendation to use unambiguous action verbs (Fry, Ketteridge et al. 2009). O'Brien's guide includes practical tables of verbs adapted from the work by Kennedy et al, recommending the use of one active verb and selecting a verb that indicates the 'level' (O' Brien 2008).

The learning outcomes were classified into those that were knowledge outcomes, skills outcomes and attitude outcomes. The level of the outcomes was assigned to each, using Bloom's Taxonomy for knowledge (Bloom and Krathwohl 1956) , Krathwohl's Taxonomy for attitude (Krathwohl, Bloom et al. 1964) and Dave's Taxonomy for psychomotor skills (Dave

1970). Miller's pyramid was used to identify the learning outcomes in the lower level of the pyramid, knows and knows how (Miller 1990), which can be assessed an objective written exam (Case and Swanson 1993).

2.3.3 Blueprint

A number of options were considered for the blueprint weighting. While the curriculum is divided into nine sections, some sections e.g. Section 1 - General Microbiology, is much larger than other sections e.g. Section 4 - Mycology. Some sections have fewer learning outcomes in the knowledge domain, e.g. Laboratory Management, which has a high proportion of learning outcomes in the behavioural (attitudes) domain (Bloom and Krathwohl 1956) and cannot be effectively assessed by a written objective test. The primary options were:

1. Give equal weight to each of the nine sections of the curriculum
2. Give more weight to larger sections of the curriculum
3. Weight sections guided by the scope of practice survey
4. Weight sections by professional consensus
5. Weight by a combination of 2 and 3.

Option 5 was chosen. Firstly, the blueprint was weighted by the number of learning outcomes in the section, then each section was adjusted to reflect the scope of practice. Areas that were within the scope of practice for >75% of countries had a weighting of 1 and areas that were within the scope of practice of 50-75% had a weighting of 0.75. There were only a few learning outcomes in the knowledge domain where the activity was reported to be within the scope of practice of <50%. These were also given a weighting of 0.75, to strike a balance between harmonisation of practice and scope of practice.

2.3.4 Test creation and delivery

Exam-Working Group

At the UEMS section of MM in 2017 an Exam Working Group was created, consisting of five volunteers from the section, one of whom would act as chairperson. The five exam-working group members were Consultant CMs from Norway, the Netherlands, the United Kingdom (UK) and Ireland. All members were experienced postgraduate educators and trainers.

The assessment literature was reviewed to consider the options for the assessment method, taking into consideration test validity, reliability, feasibility, and discrimination. Frameworks from existing UEMS European Examinations (Surgery, Cardiology, Radiology) were reviewed, to assist in creation of a new Framework for the European Examination in MM.

Question-writing and review

At the UEMS section meeting in 2018, members were asked to volunteer to become question writers. The question-writers were divided into 4 groups, each chaired by a member of the Exam Working Group, and assigned specific areas of the curriculum to consider, based on their individual areas of interest and expertise. Question-writing workshops was held, initially face-to-face and then subsequently online. The question-writing group was supported by an experienced educationalist.

Questions initially underwent an internal review within the question-writing group before being submitted for inclusion in the question bank. The questions were then reviewed by question-writers from outside their group. The questions then underwent a second, external review by a group selected from members of the UEMS section of MM. At this

stage, items were selected from this bank of peer-reviewed items for the pilot assessment and mapped to the blueprint to ensure constructive alignment with the European curriculum in MM.

The pilot exam and delivery platform

It was agreed that the exam would be delivered on an online platform. Specifications were agreed for the exam platform requirements (Appendix 3). The exam delivery platforms were assessed against the agreed criteria.

The invitation to participate in the exam was advertised in the ESCMID newsletter. ESCMID is a global network with a large trainee membership, facilitating communication on an international scale. It was also advertised on the UEMS section of MM website and through the membership of the section and their National Societies. ESCMID is the largest society for Clinical Microbiology in Europe, with 12K members and 18K attendees at the national conference.

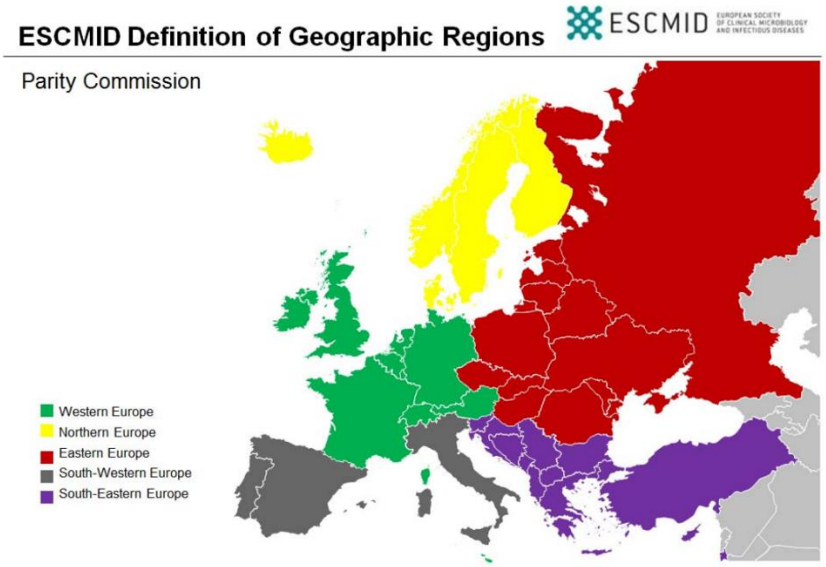
An application form was created along with a document outlining the exam terms and conditions (Appendix 4). The exam is designed to be an 'exit' exam for doctors training to be specialists in MM. The ETR for MM recommend 60 months of specialist training. Eligible candidates had to be a medical doctor with a minimum of 42 months specialist training in Clinical Microbiology.

Candidates were invited to sit this pilot examination free of charge, and informed that they would subsequently be invited to sit the first EEMM at a discounted fee. The pilot exam was delivered online in March 2021.

2.4 Results

2.4.1 Scope of Practice survey

Thirty-nine responses were received from 16 countries: Austria, Croatia, Denmark, Finland, France, Germany, Ireland, Latvia, the Netherlands, Norway, Slovakia, Slovenia, Spain, Sweden, Turkey, UK. There was representation from all of the five ESCMID-defined Geographic areas, as defined by the ESCMID Parity Commission.



The scope of Practice survey in 2019 had responses from all of the five ESCMID defined Geographic areas. The highest response was from countries in Western Europe. The lowest response was from countries in South-western Europe
The column labelled section MM indicates the number of countries in each area in the UEMS section of Medical Microbiology

Geographic Region	2019	Section MM
Western Europe	6	9
Northern Europe	4	4
Eastern Europe	3	6
South-Western Europe	1	2
South-Eastern Europe	3	6

Figure 2.2 Response to the scope of practice survey.

The scope of Practice survey in 2019 had responses from all of the five ESCMID defined Geographic areas. The highest response was from countries in Western Europe. The lowest response was from countries in South-western Europe. The column labelled section MM indicates the number of countries in each area in the UEMS section of MM.

Laboratory

Respondents were asked to detail how frequently specific roles were undertaken by the Microbiologist in 2019. The laboratory tasks are divided into sample pre-analytical phase, sample analytical phase (Table 2.1) and sample post-analytic phase and laboratory leadership duties (Table 2.2). The three laboratory roles most commonly selected as being *always* the role of the microbiologist were deciding the Standard Operating Procedures/Protocols for antimicrobial sensitivity test (n= 23, 63.8%), deciding the Standard Operating Procedures/Protocols for Identification of organisms (n= 20, 57.1%), and interpreting the test results to give clinical advice (n= 19, 52.8%). A full breakdown is presented in Table 2.1 and Table 2.2.

Table 2.1 looks at the reported frequency with which specific laboratory tasks are undertaken by the CM. The Consultant Microbiologist reported how frequently the task fell within their scope of practice as 'never', 1-25% of the time, 26-50% of the time, 51-75% of the time, 76-99% of the time or it was 'always' their role to complete the task. This table has been divided into the sample pre-analytical phase and sample analytical phase.

TASK	Reported frequency of the task being the role of the Clinical Microbiologist					
	Never	1-25%	26-50%	51%-75%	76%-99%	Always
Pre-analytical phase – handling/preparation of the sample						
Specimen set up (n= 37)	16.2% (6)	51.1% (19)	0% (0)	0% (0)	18.9% (7)	13.5% (5)
Analytical phase – performs the test						
Performs the Microscopy (n= 35)	2.9% (1)	34.3% (12)	11.4% (4)	20% (7)	20% (7)	11.4% (4)
Performs the Molecular assays (n= 32)	21.9% (7)	43.8% (14)	12.5% (4)	0% (0)	12.5% (4)	9.3% (3)
Performs the antimicrobial susceptibility tests (n= 34)	11.8% (4)	38.2% (13)	14.7% (5)	5.9% (2)	14.7% (5)	14.7% (5)
Performs the tests for identification of organisms e.g. set up API or MALDI-Tof (n= 34)	17.6% (6)	50% (17)	5.9% (2)	2.9% (1)	11.8% (4)	11.8% (4)
Performs the serology assay/test e.g. for HBsAg (n= 34)	26.5% (9)	50% (17)	5.9% (2)	5.9% (2)	2.9% (1)	8.8% (3)
Performs the examination of blood/stool other tissue for the presence of protozoa/helminths (n= 33)	9.1% (3)	48.4% (16)	6.1% (2)	12.1% (4)	9.1% (3)	15.2% (5)
Performs the tests for the examination of skin/hair/nails (n= 34)	17.6% (6)	47.1% (16)	14.7% (5)	2.9% (1)	11.8% (4)	5.9% (2)

Table 2.1: The frequency with which specific laboratory tasks (pe-analytical phase and analytical phase) are reported to be undertaken by the Clinical Microbiologists in Europe in 2019.

TASK	Reported frequency of the task being the role of the Clinical Microbiologist					
	Never	1-25%	26-50%	51%-75%	76%-99%	Always
Post-analytical phase evaluate & report the result						
Decides when susceptibility testing is required (n= 36)	0% (0)	8.3% (3)	5.6% (2)	11.1% (4)	36.1% (13)	38.9% (14)
Decisions on extra sensitivity testing (n= 35)	0% (0)	5.7% (2)	0% (0)	8.6% (3)	45.7% (16)	40% (14)
Reports and approves the FINAL Microbiology result/report (n= 36)	0% (0)	5.6% (2)	13.9% (5)	11.1% (4)	25% (9)	44.4% (16)
Interpret the test results to give clinical advice (n= 36)	0% (0)	5.6% (2)	8.3% (3)	13.9% (5)	19.4% (7)	52.8% (19)
Interpret the test result to give infection prevention and control advice (n= 36)	0% (0)	11.1% (4)	13.9% (5)	11.1% (4)	22.2% (8)	41.7% (15)
Decides when a sample or organism will be referred to a reference facility for additional testing (n= 36)	0% (0)	2.8% (1)	8.3% (3)	8.3% (3)	30.6% (11)	50% (18)
Laboratory Leadership duties						
Deciding the Standard Operating Procedures for Identification of organisms (n= 35)	0% (0)	0% (0)	2.9% (1)	8.6% (3)	31.4% (11)	57.1% (20)
Developing Standard Operating Procedures for specimen set up, culture conditions and microscopy (n= 35)	0% (0)	2.9% (1)	11.4% (4)	11.4% (4)	34.3% (12)	40% (14)
Developing the Standard Operating Procedures for serology assays (n= 35)	0% (0)	2.9% (1)	8.6% (3)	11.4% (4)	37.1% (13)	40% (14)
Developing Standard Operating Procedures for the examination of samples for the presence of protozoa/helminths (n= 33)	3% (1)	6.1% (2)	3% (1)	12.1% (4)	36.4% (12)	39.4% (13)
Deciding the Standard Operating Procedures for antimicrobial sensitivity test (n= 36)	0% (0)	2.8% (1)	0% (0)	8.3% (3)	25% (9)	63.8% (23)

Table 2.2: The frequency with which specific laboratory tasks (post-analytical phase and leadership roles) are reported to be undertaken by the Clinical Microbiologists in Europe in 2019.

Table 2.2 looks at the reported frequency with which specific laboratory tasks are undertaken by the CM. The Consultant Microbiologist reported how frequently the task fell within their scope of practice as 'never', 1-25% of the time, 26-50% of the time, 51-75% of the time, 76-99% of the time or it was 'always' their role to complete the task. The table has been divided into the sample post-analytic phase and laboratory leadership duties.

Antimicrobial Stewardship

The Antimicrobial Stewardship tasks included in the survey were those usually overseen by the Lead for Antimicrobial Stewardship. The frequency with which this was reported to be within scope of practice varied across the frequencies. The Consultant Microbiologist was more frequently responsible for developing the empiric antimicrobial choices, than for developing the surgical prophylaxis guidelines.

TASK	**Reported frequency of the task being the role of the Clinical Microbiologist**					
	Never	1-25%	26-50%	51%-75%	76%-99%	Always
Antimicrobial stewardship Leadership						
Developing guidelines/protocols for Surgical prophylaxis (n= 36)	11.1% (4)	22.2% (8)	25% (9)	8.3% (3)	16.7% (6)	16.7% (6)
Developing guidelines/protocols for empiric antimicrobial choices (n= 35)	2.9% (1)	25.7% (9)	25.7% (9)	5.7% (2)	20% (7)	20% (7)

Table 2.3: The frequency with which Antimicrobial Stewardship leadership tasks are reported to be undertaken by the Clinical Microbiologists in Europe in 2019.

The table looks at the reported frequency with which Antimicrobial Stewardship leadership tasks of developing the hospital prescribing guidelines are undertaken by the CM. The Consultant Microbiologist reported how frequently the task fell within their scope of practice as 'never', 1-25% of the time, 26-50% of the time, 51-75% of the time, 76-99% of the time or it was 'always' their role to complete the task.

Infection Prevention and Control

The Infection Prevention and Control tasks included in the survey were divided into IPC leadership in the hospital (Table 2.4) and routine or day-to-day IPC tasks and community IPC tasks (Table 2.5). The scope of practice was very varied. The IPC tasks in the hospital were reported to be the role of the CM more frequently than in the community.

TASK	Reported frequency of the task being the role of the Clinical Microbiologist					
	Never	1-25%	26-50%	51%-75%	76%-99%	Always
Infection Prevention and Control leadership in the hospital						
Leads the Infection Prevention and Control team in the Hospital. (n= 34)	14.7% (5)	14.7% (5)	17.6% (6)	11.8% (4)	17.6% (6)	23.5% (8)
Lead role in Hospital outbreaks of infection (n= 36)	8.3% (3)	19.4% (7)	11.1% (4)	16.6% (6)	30.6% (11)	13.9% (5)
Lead role in matters relating to hospital building and design (n= 34)	23.5% (8)	35.3% (12)	5.9% (2)	8.8% (3)	17.6% (6)	8.8% (3)
Lead role in matters relating to Hospital ventilation requirements e.g. Standards for Ventilation in the Operating Theatre (n= 34)	14.7% (5)	38.2% (13)	5.9% (2)	2.9% (1)	23.5% (8)	11.8% (4)
Leads on development of Guidelines/Protocols for Infection Prevention and Control in the Hospital (n= 34)	11.8% (4)	17.6% (6)	17.6% (6)	8.8% (3)	26.5% (9)	17.6% (6)
Leads on development of Guidelines/Protocols for decontamination of equipment in the hospital (n= 35)	14.3% (5)	37.1% (13)	5.7% (2)	5.7% (2)	20% (7)	17.1% (6)
Leads on development of Guidelines/Protocols for preventing transmission and Nosocomial spread in the Hospital (n= 35)	11.4% (4)	17.1% (6)	20% (7)	11.4% (4)	14.3% (5)	25.7% (9)
Review and implementation of new or revised National and International Guidelines in Infection Prevention and Control e.g. Guidelines for Management of CPE (n= 35)	5.7% (2)	20% (7)	14.3% (5)	8.6% (3)	31.4% (11)	20% (7)
Oversees surveillance of antimicrobial resistance (n= 36)	2.8% (1)	2.8% (1)	5.6% (2)	16.6% (6)	41.6% (15)	30.6% (11)

Table 2.4: The frequency with which specific Infection Prevention and Control leadership tasks are reported to be undertaken by the Clinical Microbiologists in Europe in 2019.

Table 2.4 looks at the reported frequency with which IPC leadership in the hospital tasks are undertaken by the CM. The Consultant Microbiologist reported how frequently the task fell within their scope of practice as ‘never’, 1-25% of the time, 26-50% of the time, 51-75% of the time, 76-99% of the time or it was ‘always’ their role to complete the task.

TASK	Reported frequency of the task being the role of the Clinical Microbiologist					
	Never	1-25%	26-50%	51%-75%	76%-99%	Always
Infection Prevention and Control day-to-day duties						
Gives advice on Nosocomial and community-acquired infections in which environmental factors play a role (e.g., food, water, air) (n= 34)	5.9% (2)	26.5% (9)	11.8% (4)	11.8% (4	26.5% (9)	17.1% (6)
Gives advice on patient isolation requirements (n= 36)	5.6% (2)	36.1% (13)	22.2% (8)	8.3% (3)	13.9% (5)	13.9% (5)
Gives advice on prevention of device-related infection (n= 36)	8.3% (3)	27.8% (10)	22.2% (8)	11.1% (4)	16.7% (6)	13.9% (5)
Infection Prevention and Control Leadership in the community						
Leads the Infection Prevention and Control team in the Community (n= 32)	31.2% (10)	25% (8)	21.9% (7)	3.1% (1)	9.4% (3)	9.4% (3)
Lead role in Community outbreaks of Infection (n= 33)	12.1% (4)	42.4% (14)	9.1% (3)	6.1% (2)	21.2% (7)	9.1% (3)

Table 2.5: The frequency with which other specific Infection Prevention and Control tasks are reported to be undertaken by the Clinical Microbiologists in Europe in 2019.

The table looks at the reported frequency with which specific IPC tasks, divided into routine or day-to-day IPC tasks and community IPC tasks, are undertaken by the CM. The Consultant Microbiologist reported how frequently the task fell within their scope of practice as ‘never’, 1-25% of the time, 26-50% of the time, 51-75% of the time, 76-99% of the time or it was ‘always’ their role to complete the task.

Clinical Microbiology

The Clinical Microbiology tasks were divided into patient-facing clinical review and ward rounds (Table 2.6) and MDTs and other duties (Table 2.7). The scope of practice was very varied. Performing a physical examination of a patient was not within the scope of practice of most CMs, however, providing a clinical advisory service by telephone to both hospital-based and community-based clinicians was the more frequently reported to be within the scope of practice.

TASK	Reported frequency of the task being the role of the Clinical Microbiologist					
	Never	1-25%	26-50%	51%-75%	76%-99%	Always
Patient-facing Clinical review						
Takes a clinical/infection history from the patient or patients' primary clinician regarding the patient presenting with an infection/infection syndrome (n= 32)	25% (8)	40.6% (13)	9.4% (3)	6.25% (2)	12.5% (4)	6.25% (2)
Carries out a physical examination of a patient presenting with an infection/infection syndrome (n= 32)	40.6% (13)	50% (16)	3.1% (1)	0% (0)	0% (0)	6.25% (2)
Ward rounds						
Participates in a ward round in the Critical Care Unit patients (n= 35)	14.3% (5)	37.1% (13)	11.4% (4)	8.6% (3)	8.6% (3)	17.1% (6)
Participates in a ward round of other specialist units e.g. transplant ward, burns unit (n= 34)	11.8% (4)	38.2% (13)	17.6% (6)	8.8% (3)	8.8% (3)	14.7% (5)
Participates in a ward round of non-critical care patients with a presumed infection/infection syndrome (n= 33)	21.2% (7)	42.4% (14)	9.1% (3)	12.2% (4)	6.1% (2)	9.1% (3)

Table 2.6: The frequency with which specific Clinical Microbiology patient-facing clinical review and ward round tasks are undertaken by the Clinical Microbiologist in Europe in 2019.

Table 2.6 looks at the reported frequency with which specific Clinical tasks are undertaken by the CM. The table has been divided into patient-facing clinical review and ward rounds. The Consultant Microbiologist reported how frequently the task fell within their scope of practice as 'never', 1-25% of the time, 26-50% of the time, 51-75% of the time, 76-99% of the time or it was 'always' their role to complete the task.

TASK	Reported frequency of the task being the role of the Clinical Microbiologist					
	Never	1-25%	26-50%	51%-75%	76%-99%	Always
Multi-disciplinary Team Meetings						
Participates in multidisciplinary team meeting (MDT)/ward round with specialist services e.g. Haematology, Orthopaedics (n= 35)	5.7% (2)	31.4% (11)	17.1% (6)	17.1% (6)	14.3% (5)	14.3% (5)
Other						
Advises on the diagnosis of a patient with infection of a patient presenting with an infection/infection syndrome (n= 34)	8.8% (3)	32.4% (11)	17.6% (6)	8.8% (3)	17.6% (6)	14.7% (5)
Advises on the treatment of a patient presenting with an infection/infection syndrome (n= 36)	0% (0)	38.9% (14)	16.6% (6)	8.3% (3)	19.4% (7)	16.6% (6)
Advises on the prevention of infection e.g. vaccination (n= 33)	6.1% (2)	54.5% (18)	12.2% (4)	6.1% (2)	12.2% (4)	9.1% (3)
Gives advice to clinicians by phone on diagnosis and management of infection (n= 36)	2.8% (1)	16.6% (6)	5.6% (2)	16.6% (6)	36.1% (13)	22.2% (8)
Gives advice to General Practitioners on diagnosis and management of infection (n= 35)	2.9% (1)	20% (7)	11.4% (4)	8.6% (3)	37.1% (13)	20% (7)

Table 2.7: The frequency with which specific Clinical Microbiology MDT and other tasks are undertaken by the Clinical Microbiologist in Europe in 2019.

Table 2.7 looks at the reported frequency with which specific Clinical tasks are undertaken by the CM. The table has been divided into MDTs and other duties. The Consultant Microbiologist reported how frequently the task fell within their scope of practice as 'never', 1-25% of the time, 26-50% of the time, 51-75% of the time, 76-99% of the time or it was 'always' their role to complete the task.

2.4.2 Curriculum and Blueprint

The assessment blueprint was created to align to the curriculum, and intentionally weighted to prioritise larger sections of the curriculum and to reflect findings from the scope of practice survey. Learning outcomes that were not assessable by a knowledge-based exam were of necessity excluded. The examination is an Objective test, testing in the lower level of the cognitive domain. Testing in the remaining domains will not be part of the EEMM at this time. Other assessment techniques that could be used by the training site to examine the remaining areas of the curriculum are summarised in Table 2.8.

Type of Learning Outcome	**Level of domain**	**Assessment techniques**
Cognitive (Knowledge)	Higher order	Formative assessment
	Lower order	Summative assessment e.g. Objective
Skills (Psychomotor)	Higher order	Formative assessment
	Lower order	Formative assessment
Behaviour (Attitudes)	Higher order	Self-evaluation in logbook or Portfolio
	Lower order	Observation Discussion

Table 2.8: Overview of all Learning Outcomes for a Training Programme in Clinical Microbiology

The European Core Curriculum in MM has LO in the domains of Knowledge, Skills (Psychomotor) and Attitude. This table summarises the elements of an assessment programme that could be used to assess all the domains of the curriculum at the lower and upper levels of Miller's pyramid.

Section 8: General Clinical Skills							
Module description: The purpose of this module is to provide learners with the ability to take an infection history, review laboratory results, advise on treatment and/or additional diagnostic tests and communicate recommendations to primary clinician.							
		Blueprinting against assessment					
Section		MCQ	formative	Audit	Research project	logbook	K/S/A
	Weighting						
8.1.1	Take a relevant basic clinical/infection history		x			x	**S**
8.1.2	Advise on diagnosis, treatment and prevention of common clinical problems	x	x			x	**K**
8.1.3	Explain results and clinical management plans to both clinicians and patients		x				**A**
8.1.4	Assimilate clinical, laboratory and epidemiological information and use this to differentiate between infections and other conditions	x					**K**

8.1.5	Select and interpret appropriate tests	x					**K**
8.1.6	Analyse data to determine a specific or differential diagnosis		x	x			**K**
8.1.7	Liaise effectively with clinical colleagues through regular ward visits and participation in collaborative clinical activities.		x				**A**

Table 2.9: Example of Meso-Blueprinting for the General Clinical skills section of the curriculum

The learning outcomes in section 8 .1 of the outcome-based curriculum in MM have been mapped against the test methods available; the knowledge exam, formative assessments, logbook, audit and research.

Section	Subsection	K	Weighting
1.0 GENERAL MICROBIOLOGY- ***61 L.O. with Knowledge domain (K) and weighting (wt)***	1.1 Scientific basis of clinical microbiology and immune response 1.2 Laboratory safety	18 L.O. 8 L.O.	61 (34%)
	1.3 Sterilisation and Disinfection	8 L.O.	
	1.4 Handling of Specimens 1.5 Data handling 1.6 Results reporting	10 L.O. 6 L.O. 3 L.O.	
	1.7 Microscopy	1 L.O.	
	1.8 Serologic and antigen-based techniques- 1.9 Molecular microbiology and other emerging technologies-	3 L.O. 6 L.O.	
2.0 BACTERIOLOGY **15 L.O.**	2.1 General 2.2 Culture methods	7 L.O. 8 L.O.	15 (8%)
3.0 VIROLOGY **12 L.O.**	3.0 Virology	12 L.O.	12 (7%)
4.0 MYCOLOGY **11 L.O.**	4.0 Mycology	11 L.O.	11 (6%)
5.0 PARASITOLOGY **8 L.O.**	5.0 Parasitology	8 L.O.	8 (4%)
6.0 ANTIMICROBIALS **16 L.O.**	6.0 Antimicrobials	16 L.O.	16 (9%)
7.0 INFECTION CONTROL IN HOSPITAL AND COMMUNITY **21 L.O. total (K)** ***16 L.O. (wt and K)***	7.1-7.23 Infection control	21 L.O.	16 (9%)

8.0 CLINICAL MEDICINE **30 L.O. total (K)** ***23 L.O. (wt and K)***	8.1-8.10 General clinical skills - 8.11 Major clinical syndromes/systems 8.12 Mother and Infant 8.13 Miscellaneous syndromes	6 L.O. 14 L.O. 3 L.O. 7 L.O.	23 (13%)
9.0 LABORATORY MANAGEMENT **19 L.O. (K)**	9.1 General Management 9.2 Quality Control 9.3 Accreditation- 9.4 Audit and clinical governance	6 L.O. 6 L.O. 3 L.O. 3 L.O.	19 (10%)
181 L.O.	**total**	**181**	

Table 2.10: The blueprint for the knowledge-based objective assessment.

The blueprint has been divided into the 9 sections of the curriculum. In some curriculum areas it has been divided into subsections. The number of learning outcomes (L.O.) in each section/sub-section has been included. The sections/sub-sections in yellow were given a lower weighting than the sections/subsections in green. 'Wt' is number of L.O. adjusted by weighting. The column marked 'K' indicates the number of L.O. in the Knowledge domain for that section/sub-section of the curriculum.

2.4.3 Test creation & delivery

A test was created with 99 items based on the blueprint. Initially the test was a combination of SBA and EMQs. Unfortunately, the online platform used to deliver the exam could not accommodate the EMQs, so these were modified, creating a test with 99 SBAs. The pilot examination was held online with remote invigilation in March 2021.

The online platform chosen for the question-bank and exam delivery was the European Board of Radiology (EBR) European Diploma in Radiology (EdiR) platform. The EDiR platform was secure, easy to use, cost-effective and provided the working group with easy tracking of question upload, question modifications and question topics.

2.5 Discussion

The aim of the project is to create a pan-European assessment in Clinical Microbiology. One of the challenges is that although the scope of practice of Clinical Microbiology in Europe has been evolving over the last five decades, there are still differences in the scope of practice.

In this chapter, my aim was to assess extent of the differences in the scope of practice in Europe using a survey, to change the existing European Curriculum in MM into an outcome-based curriculum, and to create a blueprint for a pilot European Exam in MM, mapped to the outcome-based curriculum, with weighting by curriculum section size and scope of practice.

2.5.1 The Scope of Practice survey

Studies in the UK undertaken 20 and almost 30 years ago, describe at least three main areas of practice for the CM: Management of infection, which includes clinical interpretation of results, Laboratory management and quality and Infection Prevention and Control (Mehtar 1995, Scott 2000). To my knowledge, this is the first time data has been collected on the scope of practice in Europe to assess whether these three main areas of practice are the role of the CM in all countries.

While much survey research has historically involved large population-based data collection (Ponto 2015), my target population, the member, associate member and observer countries of the UEMS section of MM) was small, comprising of just 27 countries. Responses were received from 16 countries (59%) of the target countries. Non-response bias is recognised as a key limitation in surveys as a research tool (Burmeister 2003), however, there were responses from all five Geographic Regions, as defined by the ESCMID

Parity Commission (Figure 2.2). The geographical spread of the responses is reassuring and should impact favourably on the non-response error (Harrison and Draugalis 1997) and sampling error (Check and Schutt 2012).

There were some limitations with respect to the survey design which may contribute to measurement error (Burgard, Bošnjak et al. 2020). In some of the questions, the language used to describe some grades of staff in the laboratory differ between countries, which may have resulted in respondent misinterpretations, and impacted on some of the data quality (Belson 1981).

The survey revealed quite varied scope of practice. The transition away from benchwork, and increasingly towards more clinical practice as reported in the UK studies (Scott 2000), was less apparent in some countries. Some areas of laboratory practice, such as leadership tasks and post-analytical tasks, including reporting the final results, are relatively harmonised, and commonly the role of the CM in most countries. However, 13.5% of respondents reported always doing the specimen set up, a routine bench-work task that would often be undertaken by a junior scientist or technician in many countries. There was a lot of variation in practice in antimicrobial stewardship leadership and clinical duties, such as participation in ward rounds and MDTs. Infection Prevention and Control (IPC), a key role for many CM, is practiced in a very limited way in some countries, particularly with respect to day-to-day issues, but leadership and guideline development in IPC was within scope of practice. Overall, there is a lot of variation between tasks undertaken and frequency of tasks undertaken, with some practitioners not undertaking roles described in the European Core curriculum.

2.5.2 The OBE curriculum

The European curriculum was taken section by section and the narrative and objectives were adapted using Bloom's Taxonomy to express the content as learning outcomes. This was not the optimal approach, as Spady says:

> "*You develop the curriculum from the outcomes you want students to demonstrate rather than writing objectives for the curriculum you already have*" (Spady 1988).

Harden likens the approach of developing an OBE curriculum to an architect designing a building, the architect's plan, which starts with a vision of the future building/structure. The preferred strategy for creating the OBE curriculum would be to start with the product, what will the future CM be like, and what knowledge, skills and behaviours that will be required to take on the role (Harden, Crosby et al. 1999). This will be the approach when the curriculum is revised later this year.

Some criticism for the OBE model is heard from those that worry that the focus on outcomes will result in a loss of creativity, a loss of education for education's sake (McKernan 1993). I would argue that once you have a safe doctor who has mastered his or her art, there is room for creativity in their professional development going forwards to select their own outcome. This view supported by others, who recognise the importance of 'the product' in medical training. One of the other challenges facing educators moving from the content model is what should remain in the curriculum, the OBE curriculum, while focussing on broader outcomes and may shift the focus away from the specific subject content of the curriculum (McNeir 1993).

The OBE curriculum describes curriculum outcomes in nine areas and has resulted in a long list of outcomes. Long lists of learning outcomes are considered difficult to manage

and hard to apply in practice, but it allowed the question-writers to focus their questions, and avoided huge changes in the curriculum that would have required a formal process of curriculum revision and subsequent ratification by council.

In future revisions of the curriculum, the process will start with the outcomes we want the trainees to demonstrate, rather than writing learning outcomes based on an existing curriculum (Spady 1988).

2.5.3 The blueprint

The results of the scope of practice survey created a challenge for the blueprint.

> "*The very nature of outcome-based medicine forces one to address inherently controversial issues*"(O'Neil 1994).

This was the case with the decision around creating one blueprint versus more than one blueprint to reflect the different practice, particularly in IPC in Europe. This is a challenge that was also faced by Clinical Neurophysiology (Cole and Kamondi 2023). Cole describes both diversity of practice and recognition of Clinical Neurophysiology as a medical mono-specialty, versus a subspeciality. There was much discussion around creating a 'lower-level curriculum' requiring fewer areas of competence, versus maintaining the aim to encourage 'high level practice' in Clinical Neurophysiology. The authors' recommendation was a modular approach to practice and examinations, with primary modules for all trainees and additional advanced modules, with flexibility for mono-specialty recognition and subspeciality recognition (Cole and Kamondi 2023).

Creating two blueprints and two exams would mean less opportunities for mobility of specialists between countries, one of the goals of harmonisation of scope of practice (EUR-Lex 2005). A survey of training in Clinical Microbiology in Europe, undertaken less than two

years later, is optimistic for harmonisation of training (Doyle, Boyle et al. 2021), predicting the likelihood of more harmonisation of practice.

In an effort to achieve a balance between harmonisation and fairness, a decision was made to create one blueprint, but to use the information from the survey to weight this blueprint, based on scope of practice. It was agreed that the European Exam would be an assessment of learning and for learning, assessing the knowledge doctors needed for practice today, while also driving harmonisation.

2.5.4 The pilot European Examination in Medical Microbiology

The test created was a knowledge-based exam, using MCQs and EMQ which test at the two lower levels of Miller's pyramid. The pilot EEMM does not examine the domains of psychomotor skills or attitudes. The Exam Working Group were aware of this limitation and see the EEMM as one part of an overall framework for training and certification as a specialist in Clinical Microbiology.

The draft framework document sets the exam within an overall training and assessment framework which includes a requirement for the trainee to participate in audit, research and a core training programme in Clinical Microbiology in order to obtain a formal qualification, such as fellowship (Figures 1 and 2 of draft Framework document in Appendix 5). The draft Framework document also recommends that trainees also undergo formative assessment during their training programme and that trainees keep a record of their training goals during training, using either a logbook or checklist, to provide evidence of completion of the UEMS section MM ETRs. The candidates applying to sit the exam required recommendations from two independent referees, who certify that the candidate is a medical doctor, is in training in Clinical Microbiology and has gained

sufficient experience to sit the exam. The exam is designed as an 'exit' exam and intended to be taken towards the end of training. The ETR for MM recommends five years (60 months) of specialist training. It was agreed, therefore, that candidates should have completed a minimum of 42 months specialist training prior to sitting the exam. The draft framework document also requires the trainee to undertake research and demonstrate proof of presentation of that research at a national/international congress/meeting.

Using a single test method, an objective test has well described limitations. It made it difficult to assess some areas of the curriculum, such as laboratory management. This is a concern, as many trainees feel that training in laboratory management is an area of weakness at the end of their training (Kass, Crawford et al. 2007). The development of a second assessment method, such as an OSPE, could address this assessment gap (Belovarac, Zabar et al. 2021).

> "*Examinations drive the educational system, because they convey in the most clear and realistic terms what students must learn or do in order to succeed*".
>
> (Miller 1990).

Therefore, creating only one curriculum and including IPC in that curriculum will make it more likely that the learners will prepare for the examination by studying IPC. While this knowledge may represent knowledge at the lower end of the Miller's pyramid, it will create a foundation on which to build on. If the doctor choses to practice in a country where IPC is part of their remit, this knowledge foundation may increase their career and mobility options.

The pilot exam used volunteer candidates to sit the exam. The candidates were not charged a fee and were given the incentive of a reduced fee in the future if the wished to

sit the real exam. The use of volunteers to pilot the exam and assess the standard of the content has potential limitations. Volunteer bias may affect the external validity and generalisability of a research study (Sackett 1979). In the context of the pilot exam, it is possible that candidates did not study for the exam, the way they might for a real exam. This may affect the assessment of the content. The candidates' survey tried to overcome this limitation by asking candidates their opinion on the exam standard and by asking candidates if they were sitting any other exit high-stakes exam at the time and how the pilot exam compared. The psychometric analysis of the exam results, which compared question difficulty and overall result (Chapter 3) was also reassuring.

The absence of a fee resulted in a good sample size in the pilot, so the use of the volunteer candidates was a good test of the exam platform's usability and the overall exam process. This was captured in the candidate survey (section 4.4.1).

Chapter 3: Psychometrics and Standard setting

3.1 Background

> "One validates, not a test, but *an interpretation of data arising from a specified procedure*", Lee Cronbach (Cronbach 1971).

The three critical elements in a high-stakes assessment are validity, reliability and defensibility (Hecker and Violato 2009). The psychometrics and statistical analysis of the test are discussed in this chapter, and I consider the interpretation of data arising from the test. Validity was introduced in Chapter 2 and will be discussed further in the context of the test results and reliability. All measurements are unreliable, whether it is a physical measurement or an educational measurement. In this chapter, I will discuss reliability i.e. the consistency of the measurement. Reliability is necessary for validity, but does not guarantee it (Hecker and Violato 2009). The psychometric criteria will determine credibility in the scores and their consequences (Norcini, Anderson et al. 2011).

In this chapter, I will also consider *defensibility*; what is important for defensibility are the standards of the assessment procedure (Hecker and Violato 2009). Standard-setting gives meaning to the test results and lies at the heart of validity (Wiliam 1996). I discuss the general concepts and approach to standard setting, type of standards and standard-setting methods. I discuss two standard-setting methods in detail, the Angoff Method (Angoff 1971) and the Bookmark Method (Lewis, Mitzel et al. 2012). The cut-score for the pilot test was calculated using both of these standard-setting methods and the results are compared and discussed. This work informed the decision around the choice of standard-setting method for the first European Exam in MM.

Throughout the chapter, I discuss a number of elements critical to quality-assurance in any assessment procedure, including reliability, validity, credibility, and generalisability. Failure to attend properly to these elements will impact on the quality of decisions we make around the test.

3.1.1 Psychometrics and Test Statistics

> "*Psychometric analysis aims to quality assure the reasonableness of the interpretation of assessment outcomes and is a major source of validity evidence*" (Coombes, Roberts et al. 2016).

As previously described, it is vital that an assessment consistently and accurately measures that which is it designed to assess (De Champlain 2010). Criterion validity (in the context of medical assessment) is the extent to which an assessment score accurately measures the construct which it is supposed to be measuring, and is often reported as a correlation coefficient (Hecker and Violato 2009). This can be further divided into *predictive* or *concurrent* criterion validity; predictive validity relates to how well the test results will predict the future performance of the candidate, whereas concurrent validity considers how the test aligns with current performance.

In any test, the test is made up of a sample of all test items that might examine the domains being examined. In theory, the sample is taken from an infinite pool of possible items. No examination is a perfect examination, and all test results will have a measurement error which is not related to the specific domains targeted in the test. Test theory is used to explain the relationship between the estimated performance and the actual test scores. The more reliable the test, the more likely it is that a similar result would be obtained if it were to be re-administered (Downing 2003, Downing 2004). Validity can,

however, be lost at the expense of reliability; an assessment repeating the same question 50 times over would have high reliability, but low validity. A psychometric analysis of a test is undertaken to improve the exam. The analysis will assess many aspects of the test including the reliability of the test, the difficulty of each item in the test, and the discrimination of each item. It may include an analysis of the distractors; it may look for possible bias and may provide a means to link multiple test forms e.g. from one year to the next. Given the varying nature of assessment types, there are various statistical methods (psychometrics) employed to test for reliability. In MCQ assessments, the items only have two outcomes, fully correct or fully incorrect, and thus are dichotomously scored items. The recommended methods for MCQs are Classical Test Theory and Item Response Theory (De Champlain 2010).

Classical Test Theory

Classical Test Theory (CTT) is a simple linear model, put forward by Charles Spearman in 1904 where only two factors are considered to influence the examinees observed score (X), that is their true score (T) and error (E), such that X = E + T. Various parameters may be calculated to assess the performance not only of individual items on the test (item statistics), but also the reliability of the test as a whole (De Champlain 2010). A major advantage of Classical Test Theory is its mathematical simplicity. While substantial computational power may be required to perform analyses for larger cohorts (particularly for test coefficients such as Cronbach's alpha), these analyses can be automated, and thus performed rapidly after the assessment is completed.

Item facility

In the past this has been described as mean score or *Item Difficulty* but is now typically referred to as Item facility. Simply put, the *p* value is the proportion of students who will get a test item correct, and a low p value indicates that the test item is difficult. The *p*-value is calculated using the formula*:*

$$P=R/T$$

where p is the item difficulty index, R is the number of correct responses to the test item and T is the Total number of responses, correct and incorrect (Sabri 2013). The p-value ranges from 0.0 to 1.00. Above 0.9 is considered easy and below 0.2 is considered a very difficult item (Sabri 2013). A good test will have test items which range from the lowest level of difficulty to the highest level of difficulty, in order to test the anticipated range of ability levels of candidates expected to sit the assessment (De Champlain 2010).

Discrimination

In a quality assessment, test items should be able to discriminate between candidates with a high ability and candidates with a lower ability, and so item discrimination values help to identify items that are not functioning correctly. For example an item that does not really fit, such as an easy item that is unexpectedly answered incorrectly by high-performing candidates, or a difficult item which is unexpectedly answered correctly by weaker candidates (Boone 2016). An item in which weaker candidates do better than the higher performing candidates is considered to be *negatively discriminating* (De Champlain 2010). The *item discrimination* (a) is the rate at which the probability of getting an item correct change with the candidates' ability level. It usually does not exceed 0.2 and generally lies between 0 and 0.2. It is useful to identify test items that are not functioning

well. A low point-biserial value e.g. <=0.2, can be used to identify items that need to be reviewed by the content experts to see if they should be excluded from the test when it is scored (De Champlain 2010). Items with a negative discrimination should be reviewed (Sabri 2013).

Point biserial or equivalent coefficient

As previously stated, MCQ items can only have two outcomes, fully correct or fully incorrect, and thus are dichotomously scored item. Thus a point *biserial* co-efficient gives an estimate of how likely it is the answer given by a candidate for an individual item in the assessment examination is reflective of, or correlates to, their overall score (De Champlain 2010). Point-biserial and biserial coefficients are commonly used discrimination indices and may be corrected or uncorrected (where the relationship between an individual item and the test score is calculated *after* removing the item score from the total test score). They provide an estimate of how well the response to one item in the test correlates with the score for the entire test (De Champlain 2010). The correlation is commonly known as the Pearson Product-moment correlation. The Pearson, *r* coefficient ranges between -1 and 1. A higher value indicates a powerful discrimination power and a highly discriminating item indicates that students with high scores got that item correct (Sabri 2013).

Test-level statistics

Classical Test Theory also provides statistics for the test as a whole, such as the Mean Score of candidates, Standard Deviation and Standard Error of the Mean (SEM). While Standard Deviation and Standard Error of the Mean appear similar, they do have different statistical inferences. Standard Deviation is the variability, or dispersion, of the *individual* data points from the mean. Standard Error of Measurement (SEM) measures how far the *sample mean of the data or cohort* is likely to be from the true population

mean, and so estimates the expected variations in score if repeated sampling occurred (De Champlain 2010). Thus, the SEM can be used to calculate confidence intervals that contain either 95% or 99% of expected observations.

Reliability

The reliability of measurement has been described as a '*signal to noise*' ratio. The mark the student achieves in the test is made up of the 'signal', the true score and the 'noise', the error (McNamara 2010). The reliability of a test equates to the reproducibility of the scores in the test, the *consistency of measurement*. It is necessary for validity but reliability does not guarantee validity (Hecker and Violato 2009). Given that there are many methods of assessment, there are also a number of methods of determining reliability, including test-retest, parallel forms, split-half, Cronbach's alpha, generalisability theory (Hecker and Violato 2009). For Multiple Choice assessments, the most commonly-reported reliability coefficients (r_{xx}) are KR-20, KR-21 (Kuder and Richardson 1937) or Cronbach's alpha (Cronbach and Meehl 1955). Cronbach's alpha, which is commonly used, was developed by Lee Cronbach in 1951 and is expressed as a number between 0 and 1. Cronbach's alpha measures how consistently candidates perform across items in the test. It is the ratio of true score variability to observed score variability. The Cronbach alpha values are often interpreted as follows; a score of < 0.5 is deemed poor, because 50% is a true measurement and 50% is error. A score of 0.5 - 0.6 is adequate, 0.61 - 0.70 is good, >0.71 is very good and a score of > 0.8 is excellent (Hecker and Violato 2009). It is generally believed that Cronbach's alpha should exceed 0.8 for high-stakes exam (which means that 80% of variation in measurement is due to consistent and stable measurements, but 20% is due to errors in measurement). A high alpha indicates that the responses for the candidate across all the test items is consistent. However, if the test length is too short,

with too few items, the value of alpha is reduced purely as a function of mathematical limitations (Tavakol and Dennick 2011). The Spearman-Brown prophecy formula (SB formula) can be used to predict the impact that the number of test items will have on the reliability co-efficient (Pamphlett 2005, Royal and Hedgpeth 2015). There is an ever-present tension between validity and reliability, however a very high alpha (high reliability) may indicate that some test items are testing the same question and are therefore redundant (low validity) (Tavakol and Dennick 2011).

So, Cronbach's alpha measures the proportion of observed variance a candidate demonstrates on different test items in a test, that is due to true score variance. However, when the test purpose is to assess Mastery, it may be more important to assess *decision consistency*, whether the passing candidate has truly achieved the standard deemed to be required for mastery. In an exam to assess mastery, the focus should be on performance and precision around the cut-score rather than the entire exam (De Champlain 2010).

The advantage of CTT is that they are simple models and easy to use without the mathematical knowledge required for IRT. CTT can be used to assess the difficulty of items. CTT does not taken into consideration the question difficulty, the item discrimination and guessing. CTT is not considered to be very good for an examination that is testing mastery or a specific ability level. When mastery is the goal, it is important to maximise reliability or precision of measurement at a specific one point on the score scale. (De Champlain 2010). In classical response theory, there is an assumption that the measurement error is the same for all scores, so it is not appropriate for tracking scores from one year to the next, unless anchor items are placed in both assessments, and specific adjustments then made on the basis of item analyses for these anchors.

Item Response Theory

Item Response Theory (IRT) overcomes many of the shortfalls of CTT. An IRT analysis of assessment takes into account both the student and the exam item behaviour (Tavakol and Dennick 2013). Although it has its origins in Louis Leon Thurstone's work on the theory of intelligence in the 1920s, the seminal works were those of Lord (Lord 1952) and Rasch (Rasch 1960). In IRT, mathematical models are used which assume that the candidate's ability and item difficulty determines the *probability* that the candidate will answer a question correctly. It assesses to what extent the actual score the candidate gets in a test reflects their true knowledge (De Champlain 2010). It is used to estimate the *probability of a correct response* on an item in a test, as a function of the item's difficulty and discrimination, and the ability of the person taking the test in terms of the knowledge or skills that are being assessed. An underlying assumption in IRT is *item invariance*. Item invariance is where the estimated parameters of items, e.g. item difficulty, remain consistent across different groups or conditions.

The simplest form of an IRT model is the "1-Parameter logistic model" (1-PL), which predicts the *probability of a candidate giving the correct response*, given the difficulty of the test item and the performance of the candidate (Figure 3.1). An Item Characteristic Curve (ICC) can be created for each item in the test. In the 1-PL model, the item difficulty parameter is the natural log of the odds of success on e.g. a dichotomous item (correct/incorrect), and the discrimination factor is fixed for all items. The ability estimate values are on the x-axis, the probability of a correct response is on the y-axis. When the probability of the correct response is 0.5, the ability estimate at that point corresponds to the items difficulty (De Champlain 2010). Using *Mean Square Statistics (MNSQ*), the test item fit can be assessed to identify misfit items that may need to be reviewed. A misfit

item is one that does not contribute to useful measurement (Boone 2016). It is also possible to review 'person-fit statistics which will identify examinees whose pattern of response is unusual (Boone 2016). In the context of high-stakes assessment, this may raise suspicion of guessing or cheating (Felt, Castaneda et al. 2017).

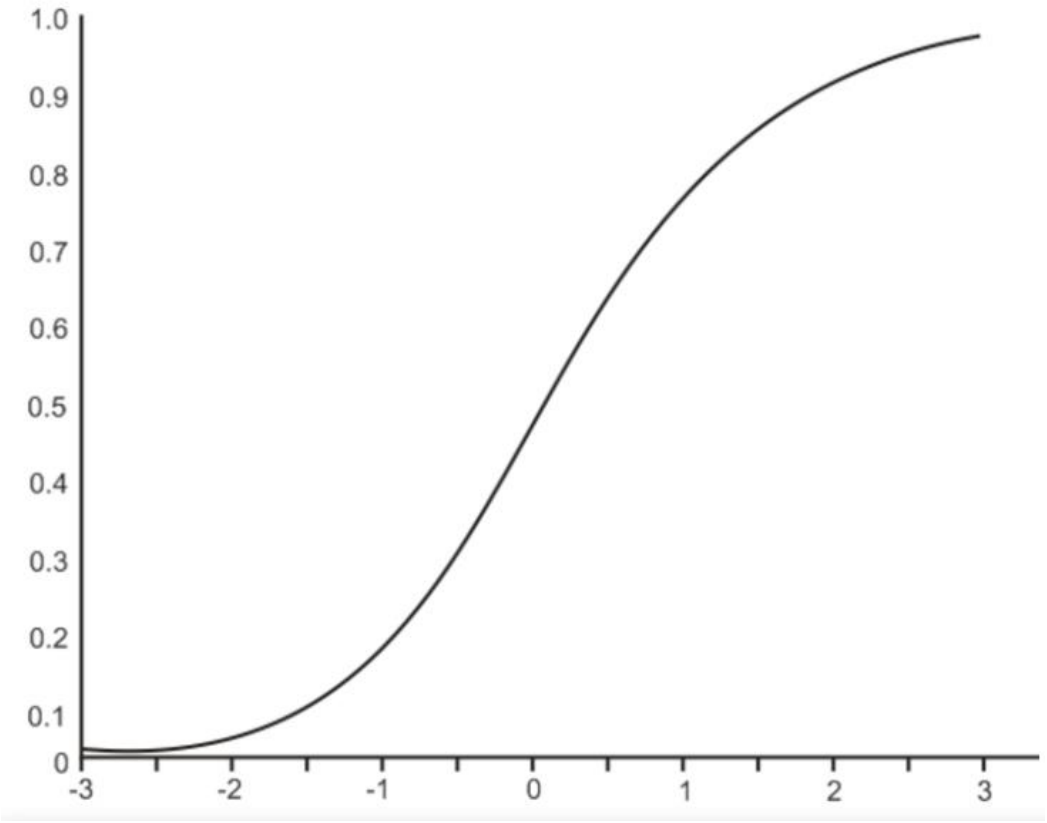


Figure 3.1 One parameter Logistic Model (1-PL) curve

Graph showing 1-PL curve, image reproduced from "The One-Parameter Logistic Model (1PLM) And Its Application in Test Development" (Umobong 2017).

Later, in 1960, a Danish mathematician developed the Rasch model (Rasch 1960), succinctly described by Boone as how we "*consider the difficulty of a test item along a variable and the overall ability level of the test taker with respect to the variable*" (Boone 2016). The Rasch model is similar to the 1-PL model, but the Rasch model constrains the Item Discrimination (a) to 1, whereas as previously described the 1-PL model does not

limit the discrimination factor to 1, and more commonly *"modelled to be parallel with a slope of 1.7 (approximating the slope of the cumulative normal ogive)"* (Linacre 2005). The Rasch Model is considered by many to be the superior model because it focusses on the test items in terms of one parameter only, the item difficulty. A further advantage is the minimum sample size required for analysis; while a 1-PL model is estimated to require a minimum sample of 200 (Downing 2003, De Champlain 2010), a Rasch model may be performed with as few as 20 (Linacre 2005). In the Rasch model, the item difficulty and student ability need to be measured in the same units. The units are called 'logits' (Tavakol and Dennick 2013). High logits (positive) suggest a higher level of ability, lower logits (negative), suggest lower ability. Using logits, the student ability and item difficulty can be displayed on the same scale (Tavakol and Dennick 2013).

The 2-parameter logistic model (2-PL) is more mathematically complex and considers two parameters, the item difficulty and the item discrimination (Figure 3.2). While useful, this 2-PL model requires large cohort sizes (>500 candidates and 30 test items) and so is only feasible for large test centres (Hulin, Drasgow et al. 1982, Downing 2003, De Champlain 2010). The three parameter model (3-PL) adds a further 'guessing' parameter (pseudo-chance) to item difficulty and item discrimination, but requires even larger sample sizes (>1,000 candidates and 60 test items) for highly accurate information (Hulin, Drasgow et al. 1982, Downing 2003).

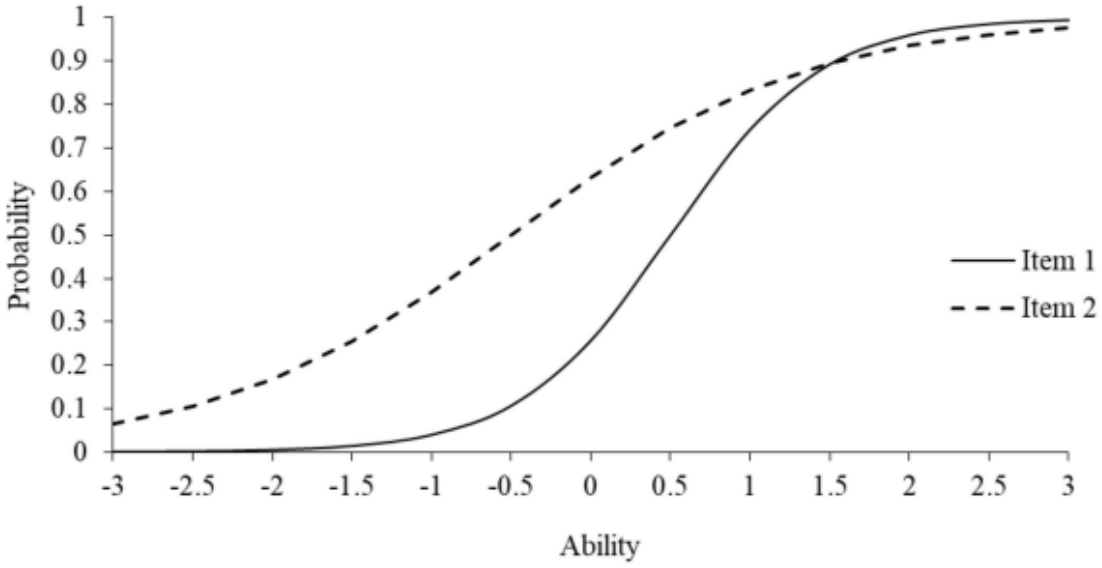


Figure 3.2 Two-parameter logistic model (2-PL) curve.

2-parameter logistic model (2-PL) curve, reproduced from Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research (Immekus, Snyder et al. 2019)

De Champlain summarises the advantage of the IRT model over the CTT model: the assumption of Item parameter invariance in IRT infers that the sample of examinees actually sitting an iteration of the test does not affect the item parameter estimates, whereas the item statistics derived through CTT are sample-dependent (De Champlain 2010).

> *"If the assumptions of common IRT models are met with the dataset, then item parameter estimates are independent of the particular sample of examinees drawn from the population of examinees for whom the test is intended (item parameter invariance property). Further, an examinee's estimated ability is not dependent upon the particular sample of test items chosen from the calibrated pool of items (ability parameter invariance property)" (De Champlain 2010).*

It may be argued that these differences between CTT & IRT may not be particularly relevant where successive samples of test items are reasonably representative, and where the ability of successive cohorts of students are equivalent, and strategies have been proposed to overcome these limitations (Bhakta, Tennant et al. 2005). However, an intuitive example may be proposed where the average ability of a sub cohort of students in repeat assessments will be lower than that of the full cohort in the original sitting, and that the item difficulty, as measured by CTT and calculated using the responses of that specific cohort, will therefore be reduced. When measured by IRT however, the difficulty of the items is factored in during scoring when estimating the probability of a candidate giving the correct response, and so these assumptions of Item invariance and ability invariance mean that exams can be compared and tracked in terms of the exam difficulty and the candidate ability (De Champlain 2010).

As previously stated, most appropriately designed assessments will have test items which range from the lowest level of difficulty to the highest level of difficulty, in order to test candidates at many levels of ability. However, because this test purpose is to assess Mastery, it may be more important to assess decision consistency, whether the passing candidate has truly achieved the standard deemed to be required for mastery. In an exam to assess mastery, the focus should be on item performance and precision around the intended cut-score, rather a wider range of ability (De Champlain 2010).

In IRT, a measurement error for all abilities and for the cut-score is estimated. For each ability value, a separate Standard Error (SE) is calculated. Therefore, we can measure how reliably we are measuring the passing score in an exam which makes decisions on mastery.

Item anchors

In an IRT framework, when running repeated assessments (for example year on year), some test items from a previously analysed assessment may be used in subsequent iterations so as to facilitate the creation of a common proficiency scale. The use of *item anchors* as a reference point to link two or more test items onto a common scale, is also referred to as linking. When item anchors are used, different forms of a test will contain a few items that are the same in each test. The proportion of items that are linked should not exceed 20% (Kolen and Brennan 2014, Fischer, Rohm et al. 2021).

3.1.2 Standard setting

When setting an assessment, the purpose is generally to measure the ability or performance of candidates, to ascertain whether they meet a required standard. Perhaps the oldest described example of a High-stakes assessment was performed by the Gilead Guards as long ago as 2000 BC. The living bible, Judges 12:5-6, describes fugitives trying to cross the river, Jordan. The fugitives' progress was halted by the Gilead guards, who asked them a question. If the answer was not satisfactory, the fugitive was killed (Cizek 2001).

Nowadays, while a method such as that described above would no doubt have Ethics committees aghast (along with the Criminal Justice System and Society at large), Standard Setting is a critical part of test development, and should be well supported by an objective evidence-base (Bejar 2008). The traditional approach to assessment where a cut-off point is defined as 50% or 70% to separate the competent from not competent has no valid evidence to support the process of determining the cut-score. This issue becomes more critical when the assessment or qualification is required for the candidate's career

progression (Friedman Ben-David 2000). In high-stakes assessment, the credibility of the threshold used to indicate whether or not the candidate's performance has met the required standard is essential for the credibility of the exam. The threshold at which the candidate is deemed to have met the standard is called the 'performance standard'. The process by which the performance standard is devised is called 'standard setting'. During the process of standard setting, positions on a 'test score scale' are determined which indicate the agreed performance standard, 'the cut scores' (Bandaranayake 2008, De Champlain 2018).

Kane, referring to the 'cut-score' as the 'passing score' in his seminal paper, explains the concept and its importance for assessment validation (Kane 1994).

> "*Validation then consists of a demonstration that the proposed passing score can be interpreted as representing an appropriate standard. The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version of the desired level of competence*' (Kane 1994)

In the 1960s, research into assessment methods saw a focus on standardised and objective assessment in an effort to move away from practices which were deemed unreliable, and even biased (Schuwirth and van der Vleuten 2020). In an assessment, there is no perfect 'golden' performance standard (Kane 1998), there is no 'perfect' passing score and Kane concludes that the best we can do is:

> '*to assemble evidence showing that the passing score and its associated performance standard are not unreasonable*' (Kane 1994)

Agreement around the performance standard has become increasingly important and consequential, and so this evidence is assembled through the process of "standard setting". Given that there are different assessment types, there are also many different standard setting processes described and reviewing them all is beyond the scope of this work, but the more common standard-setting processes in current use are well described elsewhere (Hambleton 2001, Hambleton and Pitoniak 2006, Reckase 2006, Cizek and Bunch 2007, Bandaranayake 2008, De Champlain 2018).

It is important for test validation that robust and transparent processes are followed, so that a quality, equitable and credible standard can be set. Bejar asks the question 'why is standard setting important?' in his article; he then answers that the use of rigorous standard setting methods and processes is critical, particularly in a high-stakes examination, to ensure that the test results are defendable (Bejar 2008). Bejar also emphasises the importance of ensuring that the test content and difficulty level is appropriate to the cohort being assessed, and that the question-writers understand the difficulty level required. This is essential if the cut-scores realised are to be psychometrically sound and appropriately divide the candidates into passing or not-passing (Bejar 2008). It has also been proposed that "*Standard setting be made an integral part of planning for test development*", although this can have implications for resources, primarily faculty training and availability (Cizek and Bunch 2007).

While many different methods have been described to arrive at a performance standard, the principle of following "due process" to achieve a standard that is credible and defensible should be common to them all. The performance standard has a number of components. The first is the performance level, this is where labels are created for different levels of candidate performance e.g. basic / proficient / advanced, or competent

/ not competent. The second is the performance description; this is a description of the characteristics that a candidate will have to reach a particular performance level (Hansche 1998).

Criterion-referenced or Norm-referenced

In a summative assessment, a decision must be made about the performance standard, the score at which the student or candidate will pass or fail. The performance standard can be set using a framework that is either criterion-referenced or norm-referenced (Bandaranayake 2008, De Champlain 2018). These two main categories of standard setting methods may be described as follows: (1) in criterion-referenced or absolute methods, standard setting is independent of test results and based on an agreed performance standard (Angoff 1971, Ebel 1979); (2) in norm-referenced or relative methods, standard setting is based on test results, and often determined by analysing how one candidates performance relates to another candidates performance (Norcini & Guille 2002; Downing et al. 2003).

The traditional, fixed percentage model was arguably a criterion-based method, using a pre-fixed absolute cut-off score. In the European continent this was commonly 55% or 60% correct answer, after correction for guessing. In the UK, the cut-score of 50% was commonly used. Thereafter, the top 10% of the cohort might be awarded a First Class honour; this would be an example of a relative standard, and is arguably independent of the test content and how much of the content the examinees know (Livingston and Zieky 1982). When a norm-referencing framework is used to interpret the scores, one candidate's score is compared to another, the candidates are ranked according to their score and a specified number or percentage of top-scoring candidates will pass irrespective of their score. If the results are sorted from highest to lowest, they will usually

achieve a 'bell-shaped' curve or Gaussian distribution, a continuous probability distribution that is symmetrical around its mean, also referred to as a 'normal' distribution. The concern with this interpretation is that if the group of candidates are very weak, the top/passing number/percentage of students may not have performed very well. As the saying goes, *you don't have to run faster than a bear, just faster than the person next to you…* The variations in test difficulty can cause significant variations in pass/fail rates. This is a concern if the exam is a licensing exam.

A more defensible approach for a high-stakes exam is to set a minimum standard that candidates are expected to be reached (criterion) in order to be successful in the exam. This criterion-referenced framework typically requires the minimum or passing standard to be agreed prior to the test, although methods such as Angoff (Angoff 1971) may be performed in retrospect. While setting an absolute cut-score, say of 50%, addresses the issue of variation in the candidates' ability, it does not address potential variation in the test difficulty between one sitting of the test and the next, and so Standards should ideally be set for each test and for each item in the test.

Norm-referenced

Maintaining credibility is essential in high-stakes assessment, so the standard-setting method needs to take the test difficulty into consideration, therefore, norm-referenced assessment is preferable to pre-fixed standards (Cohen-Schotanus and van der Vleuten 2010).

However, norm-referenced assessment, while providing information on a candidate's performance relative to the performance of their peers, does not estimate the actual level of performance of each candidate. The principle responsibility in specialist medical training is to train a competent specialist, not to place them in order of ranking (Turnbull

1989). A *compromise method*, as a practical and affordable alternative to criterion-referenced has been described by Cohen-Schotanus & van der Vleuten. In this method, the high-performing candidates are used as a relative point of reference to create a pre-fixed cut-off score (Cohen-Schotanus and Van der Vleuten 1996). This is known as 'Cohen's method' and is based on the principle that the best-performing students are one stable factor, even when test difficulty fluctuates. The authors compared setting the cut-off score at 60% of the highest score with 60% of the 95th percentile. The 95th percentile was deemed the best point of reference, resulting in less fluctuations in the failure rates. The authors concluded that while standard-setting panels are preferred for high-stakes assessment, for inhouse tests, the compromise method was both acceptable and affordable (Cohen-Schotanus and van der Vleuten 2010).

Criterion referenced; Examinee-centred or test-centred

Criterion-referenced standard setting methods can be categorised as being 'examinee-centred' or 'test-centred (Downing and Yudkowsky 2009). In an examinee-centred method, such as the Borderline group method, two pieces of information are needed, the examinees test scores and a global judgement on the overall ability level of all candidates in order to set the standard, after the examination has been completed (Livingston and Zieky 1982). The judges or examiners are asked to think about the characteristics they would associate with the borderline, or barely-passing examinee and then to think about specific examinees that fit these characteristics or standard (Livingston and Zieky 1982). It is essential that the judges are qualified to make these judgments and so should be aware of the standard that an examinee should have to pass at the time of the test, giving a global rating to each candidate as to whether they are pass, borderline or fail, in addition to granular and accurate scoring of each criterion (Livingston and Zieky 1982). After the

assessment has been taken, the median score from the candidates that those examiners have identified as borderline is used as the cut-score for the test (Livingston and Zieky 1982). The median is used rather than the mean, because it will be less affected by outliers or very high/low scores. This is important, because examinees with very high/low scores, likely did not belong in the 'borderline-group' after all. The borderline group method is 'group-dependent' and based on the likelihood that the examinees and judges as a group, represent the population. This is necessary for the cut-score to be considered credible (Livingston and Zieky 1982). If the borderline method is working well, the test scores of the borderline group should be clustered close together. The borderline group method is easy to use, but has a number of disadvantages; firstly it is time-consuming and secondly, it requires a large group and a major limitation of this method is that the judges must know the examinees and the required standard, so unfortunately this is not a suitable method for a European assessment, where the judges are likely to know few if any of the candidates.

In test-centred standard setting methods, judgements are made about the items in the test. The judges will consider a hypothetical examinee, whose ability is such that they are just meeting the required performance standard, and then consider how this hypothetical examinee might be expected to perform on an each test item (Hambleton and Pitoniak 2006). A number of test-centred standard setting methods have been described including the Angoff method (Angoff 1971), the Ebel method (Ebel 1972), the Nedelsky procedure (Nedelsky 1954), Jaegers method (Jaeger 1982) and the Bookmark method (Lewis, Mitzel et al. 1996). Two of the most commonly used test-centred methods are the Angoff method and the Bookmark method (Friedman Ben-David 2000, Yousefi Afrashteh 2021).

Angoff Method

The Angoff method is one of the most commonly used standard-setting methods worldwide (Cizek 1996). In the Angoff method, the judges spend some time considering and describing the characteristics of an examinee who is 'borderline', that is, the 'barely passing' candidate. Then, all individual judges are asked to individually estimate the probability that the 'borderline' candidates will answer each item on the test correctly (Angoff 1971, Downing and Yudkowsky 2009). Thereafter a period of discussion ensues, where the judges are made aware of the other judges' estimates and have an opportunity to discuss these estimates and change their own if they wish. This step is not routinely performed for all questions in the 'modified Angoff method, and instead discussion is only required to take place if excessive variance in estimates is observed (Norcini 2003). The estimated probabilities for each of the judges are added together and the average (mean) is calculated; this is then the cut-score or 'passing standard' for the test. A simplified example of this process is provided in Figure 3.3.

Question	Judge 1	Judge 2	Judge 3	Judge 4	Average
1	0.60	0.65	0.70	0.65	0.65
2	0.40	0.30	0.40	0.50	0.4
3	0.30	0.70	0.50	0.50	0.5
4	0.50	0.40	0.50	0.60	0.5
5	0.65	0.70	0.75	0.70	0.7
					2.75

$$\text{Pass mark} = \frac{2.75}{5} \times 100 \longrightarrow 55\ \%$$

Figure 3.3 The Angoff Method of standard setting

Each of the judges estimate the probability that the 'borderline' candidates will answer each item on the test correctly. The average of the judge's probability for each question is calculated. The estimated probabilities for each of the judges are added together and the average (mean) is calculated; this is then the pass-mark or cut-score.

The advantage of the Angoff method is that it is mathematically straightforward to administrate, provided one has sufficient judges, and it can be administrated before the test has been taken (Kane 1998). In Angoff methodology, the cut-score is compensatory, that is, a high score on one test item will balance a low score on another test item (Hambleton and Pitoniak 2006). Several disadvantages are described, one of the most common is the challenge for the judges in understanding the concept of the 'borderline' examinee (Elfaki and Salih 2018, Yousefi Afrashteh 2021). There is also a tendency for the judges to over-estimate the performance on easy items and underestimate on the more difficult items (Hambleton and Pitoniak 2006). The Angoff method is also criticised for being very time-consuming for the judges.

The Bookmark Method

The bookmark standard setting method was designed to overcome some of the challenges with other test-centred methods. Different types of test format are suited to different standard setting methods. While Angoff is used widely and successfully for Multiple Choice Question test format, for tests of performance evaluation or constructed response, it is less suitable. Bookmark can be used successfully for Multiple Choice Question test format and for tests of performance evaluation or constructed response (Hambleton 2001, Reckase 2006). The bookmark method combines the expert decision from the judges with measurement models to determine the cut-score. Before starting the bookmark standard-setting method, the test results are analysed, and the item difficulties are calculated using

Item Response Theory (IRT). The value for the item difficulty (p) is used to place the test items in order of difficulty in an Item Ordered Booklet (IOB). The IOB can be used to in mixed method tests e.g. a test with a combination of MCQ and constructed response items (Çetin and Gelbal 2013).

At the start of the bookmark standard setting session, the judges are then asked to think about the 'borderline', or barely passing candidate. The judges are then a sked to consider the response probability for the borderline candidate. The response probability (RP) is usually set at 0.67, this is the probability of the borderline candidate having a 67% chance of answering an item correctly (Huynh 2006, çetin and Gelbal 2013).

Starting with the easiest item on the OIB, the judges are asked to determine for each item whether in a group of borderline candidates, it is likely that there is a 67% or greater chance of them answering the item correctly. If the answer is 'yes', the judges move to the next item. The judges continue to move through the items in the IOB until they reach an item where the answer is "no". A 'bookmark' is placed at the first test item where the judges believe that a borderline candidate has a less than 67% probability of answering the test item correctly. The judges then share and discuss where they have placed their bookmarks and why they placed it there. The judges can chose to revise their bookmark (çetin and Gelbal 2013). Using IRT, the cut-score calculations are made based on the bookmark placings.

One of the advantages of bookmark is that the judges have fewer judgements to make, because the items are ordered in order of difficulty. This reduces the work for the judges, particularly when more than one cut-score is required to separate levels of proficiency and is less time consuming than Angoff for the experts determining the cut-scores.

In bookmark, the fewer judgements required may make the number of item difficulty estimation errors less (Buckendahl, Smith et al. 2002). Because bookmark is an IRT-based method, it also has the advantages of the psychometric perspective (Cizek and Bunch 2007). This method can be useful in tests which have more than one question format e.g. a test with both MCQs and questions requiring structured answers (çetin and Gelbal 2013). There may, however, be some disagreement among judges regarding the placement of the bookmark and even the order of the items (Plake 2005). There are a number of factors that might contribute to this. The items in the IOB are ordered according to the difficulty score based on the actual candidates' responses in the test. The items in the IOB are not spaced apart equally, i.e. equally distributed throughout the booklet, so there may be a cluster of items together, with similar difficulty. This may result in a difference of opinion among judges around the cluster.

A limitation of bookmark is the requirement for input from an expert in psychometrics. Buckendahl et al, described a modification of Bookmark, where the OIB is ordered using Classical Test Theory *p*-values (proportion of examinees that answered an item correctly) rather than the *b*-values (a difficulty parameter, the ability required for an examinee to have a 50% chance of getting the item correct) recommended in IRT. This may make the bookmark procedure more accessible to groups without the expertise of an experienced psychometrician (Buckendahl, Smith et al. 2002).

One of the challenges with the bookmark method is again the concept of the borderline candidate. It has been noted that asking the judges to picture a particular trainee that they had experience of who was borderline, may improve the judges understanding of this concept (Downing, Tekian et al. 2006).

Another challenge for the judges can be the Response Probability (RP) and application of the RP. It may help the judges to grasp this somewhat abstract concept if they consider a group of borderline candidates and for RP67, whether 2/3 (67%) of them might get the individual test item correct. It is important that judges are clear that the borderline candidate is the 'barely passing' candidate and not an examinee of 'average' ability.

Response Probability

The RP value determines the place on the OIB that the probability of a barely passing candidate answering the test item correctly drops below 67% (for RP67). While theoretically, the choice of RP should not affect the cut-score, it has been demonstrated that the cut-scores may differ depending on the RP value used. Usually, an RP value of 67 is used. This is because the aim is to separate the items the barely proficient candidate has mastered from those that it has not. Mitzel proposed that that an RP of 67 is required for mastery, rather than an RP of 50 (Mitzel, Lewis et al. 2001).

Several studies have focussed on the effect of different RP values. These have shown that the cut-score for RP67 were higher than those for RP50 (çetin and Gelbal 2013). It may be that the judges will progress through the OIB to later questions when at RP67 compared to RP50, creating a higher cut-score with RP67 (Hambleton and Pitoniak 2006). Brevetas looked at five different IRT models and three different RP values (0.5, 0.67, 0.8) and found that different permutations affected the bookmark difficulty location (BDL) (Beretvas 2004).

Research has been performed comparing the judges' opinion of item answering probabilities to actual item difficulty and in some studies, it has been observed that it was only moderate (çetin and Gelbal 2013). Some report that a moderate correlation is adequate (Reckase, Feuer et al. 2001), while others report that a high level of correlation

is required (Hambleton 2001). It has been suggested that the reason why the correlation may be moderate, rather than high, is because judges generally predict the RP for difficult items to be more difficult than they are and the RP for easier items, to be lower than they are (Clauser, Harik et al. 2009).

There is no clear solution to selecting the RP. One approach might be to look at the IOB and ensure that there are enough items in the area the cut-score is placed. The RP value chosen should be determined by the density of items in the potential locations (Karantonis and Sireci 2006).

Quality assurance in standard setting

Downing *et al* believe that all decisions regarding standard setting methods are policy decisions and "*there is no 'gold standard' for a passing score*" (Downing 2003). The important thing is the ***process*** of standard setting, for which four criteria should be considered; it should be systematic, reproducible, absolute and unbiased (Downing 2003).

Validity

As discussed in Chapter 2, validity is important for quality assurance in assessment. Kane proposed that three types of validity evidence be considered; procedural, internal and external (Kane 1994). Procedural evidence may be summarised as how well the standard-setting methodology was adhered to during the process, to what degree it was systematic and thus defensible in its execution. Facets to consider in the procedural evaluation are the selection of the judges, the judges understanding of the procedure, the implementation of the method and the confidence the participants have in their cut-scores (Freunberger, Luger-Bazinger et al. 2013).

The second aspect is the internal evidence; this relates to the data generated during the standard-setting process, and in particular considers the consistency of the results. When it is demonstrated that the inter-judge consistency is high, with low standard deviations between the judges, this increases the confidence in the validity of the 'cut-score' (Hambleton and Pitoniak 2006). An analysis of the variance between judges scores, such as *inter-rater reliability* or a *generalisability analysis* (Tiffin-Richards, Pant et al. 2013), can be used to evaluate the reliability of the process. The third type of validity evidence is external evidence. As the name suggests, this is based on information from other sources, such as triangulating with information from a second standard setting method, or perhaps information from the candidate's tutor or teacher (Sireci 2009).

Perhaps the final consideration for validity is how the cut-scores will be interpreted, the effect of the standard setting on the candidates results, whether the subsequent passing level is appropriate to the type of assessment (e.g. in a high-stakes assessment), and the reporting of the cut-scores to the relevant stakeholders. This is termed the *consequential validity* (Pant, Rupp et al. 2009).

Credibility

In high-stakes assessment, psychometric criteria are essential to determine credibility in the test scores (Norcini, Anderson et al. 2011). To gain credibility, it is not the method that is important, but whether that method achieves certain criteria (Norcini and Shea 1997). The method should produce standards consistent with the purpose of the test. The method should rely on well informed expert judgement. This view is supported by Downing (Downing and Yudkowsky 2009). The standard-setting process should demonstrate due diligence, should be easy to understand, explain, implement and finally, should be supported by research (Norcini and Shea 1997).

Generalisability theory

Generalisability is an approach to estimating reliability and variance in error in a situation where more than one source of error measurement is introduced (Shavelson and Webb 1991). A generalisability coefficient may also be used as a simple reliability coefficient. An estimate of reproducibility of measurement is made on a scale of 0-1.0, where 0.8 is considered a minimum requirement for a valid measurement (Wass, Van der Vleuten et al. 2001). In criterion-referenced tests, the sampling of items in the test is as important as the source of variation (Brennan 1984).

The Standard Error of Measurement (SEM)

An important consideration in analysis of test results is the Standard Error of Measurement (SEM), a measure of precision, which can create confidence bands around observed scores (Downing 2004). It is recommended by Downing (Downing 2004) and by Norcini (Norcini 1999), that we convert our estimates of reliability to the SEM. Some authors consider the SEM to be a more consistent measure of reliability than a reliability co-efficient, such as Cronbach's alpha, which may be artificially inflated if the standard deviation of marks increases (Tighe, McManus et al. 2010). The SEM establishes the confidence interval within which we would expect the candidates true score to be. SEM is the expected measurement of error or variation in scores that would occur with repeated sampling e.g. if a candidate took different tests (Ricketts 2009) (Musselwhite Thompson and Wesolowski 2018). The smaller the SEM is, the more precise the measurement capacity of the test.

The formula used to calculate SEM is different in criterion-referenced versus norm-referenced tests. The standard deviation (SD) of examinees test scores is not required to estimate the SEM in criterion-referenced, therefore, the SEM can be estimated from a test

score for a single candidate. This makes sense, because in criterion-referenced framework, we are judging the mastery of the individual candidate and not how they compare to their peers (Ricketts 2009).

Faculty Engagement in Standard Setting processes

The importance of providing training for the standard-setting judges is well described (Hambleton and Pitoniak 2006). Depending on the judges own scope of practice and areas of specialisation, they may not agree on what should be a core knowledge or skill, these local curricular differences may contribute to item disordinality (Skaggs and Tessema 2001). It is possible that this could occur in the setting of the European exam where scope of practice differs somewhat from one country to the next and therefore possibly from one judge to the next. In this situation, it is important to consider the curriculum and test construction.

During the process of reviewing items for standard setting, the judges may identify items that are weak or inappropriate. This adds to the quality-assurance of the standard setting process (Friedman Ben-David 2000). Some groups will create a Performance Level Description (PLD) through group discussion to define the 'borderline' candidate and help the judges apply the method more accurately (Yim and Shin 2020).

Performance level descriptors. Describing the borderline candidate

Performance level descriptors are used to describe the specific level of knowledge and skill that is required to achieve a passing score. They should be written before standard setting and some believe should be used in developing the test blueprint (Perie 2008). The PLD may be quite detailed in its list of expected performance standards and candidates ability, (Yim and Shin 2020). This approach has been demonstrated to work well for improving inter-rater reliability (Maurer and Alexander 1992). In Yim's paper exploring the use of

Angoff to set a standard on mock exams in nursing, she concluded that the Performance Level Description (PLD) of a candidate with the minimum level of competence, was a necessary part of the process and recommended that the PLD should be informed with information gathered from experts through conferences and research associations and include opinions from more stakeholders (Yim and Shin 2020). Alternatively, the raters may be asked to define the borderline candidate independently, but this may reduce inter-rater reliability (Fehrmann, Woehr et al. 1991).

The pilot exam in Medical Microbiology

As part of our examination simulation, we used two standard setting methods in the pilot examination to compare the experience, results and candidate outcomes.

3.2 Aims

The aim was to use the examination simulation, the pilot European exam in MM, to test the processes involved in standard setting a high-stakes exam:

To test the process of selection and training of the judges to do the standard setting.

To use psychometric analysis to assess the quality of the pilot exam.

To compare the cut-scores from two standard setting processes and the implications of the two standard-setting methods for the candidates, i.e. The percentage of candidates achieving a score above the cut-score.

3.3 Methods

3.3.1 Selection and training of the judges

Selection of judges

- The judges were selected from the members of the UEMS section of MM. The selection criteria were as follows:
    - Consultant Microbiologists
    - Significant experience in postgraduate training, at national level, in their country.
    - Experience in European postgraduate education as their country representatives on the UEMS section of MM.
    - General Microbiologists, with different sub-specialist interests e.g. laboratory diagnostics, quality, infection prevention and control, antimicrobial stewardship.
    - A mixture of additional experience/qualifications were desirable e.g. professors, an interest in quality and standards, a postgraduate qualification in medical education, experience in delivery of a national exam
    - The judges should have an age range
    - The judges should be practicing in different countries.

Training of the judges

The judges had a training session with an experienced medical educationalist.

The process started with discussion around the concept of the borderline candidate. The concept of the minimally competent or borderline candidate is a concept that many judges will find difficult to fully understand (Norcini 2003). In our study, with the guidance of an educationalist, we discussed and defined the borderline candidate during the training session. This is a common approach and works well (Fehrmann, Woehr et al. 1991). This will be discussed in more detail for each method below.

3.3.2 Assessment delivery & psychometrics

Assessment delivery

- In March 2021, the pilot examination in MM exam was delivered online with remote invigilation. Candidates were asked to check their equipment at least 24 hours before the exam (computer, internet).
- Instructions were issued to candidates explaining that the exam would be proctored by live remote invigilation and outlining the rules for exam conduct, including information on unauthorised materials e.g. smart watches, headphones, mobile phones.
- The exam proctoring was done by a combination of the members of the Exam Working Group and by experienced staff from the exam platform delivery team. The process started with a number of checks for the candidate identification and checks of the room in which they were sitting the exam.

- Candidates were made aware that they would be monitored during the exam and that the invigilator would contact them, by speaking to them, if they had any concerns or questions regarding their patterns of behaviour. Candidates were aware that they were not permitted to leave the room once the exam started.
- Each invigilator supervised 6-8 candidates. During invigilation, the invigilators communicated with each other through a separate online meeting with the platform delivery team.
- Candidates were given 150 minutes to complete the assessment.

Psychometric analysis

- Psychometric analysis was done by an experienced psychometrician. Rasch analysis was used because it focuses on item difficulty and is acceptable in smaller sample sizes. Core Rasch analysis was done using Winsteps software. Data was further analysed using R statistics Rasch analysis software.
- The candidate performance was analysed to determine the mean score, maximum score and minimum score. Reliability was measured using Cronbach's alpha and standard Error of Measurement (SEM).
- Item analysis was undertaken to assess item difficulty, discrimination, infit and outfit and distractor analysis.

3.3.3 Standard Setting - Angoff

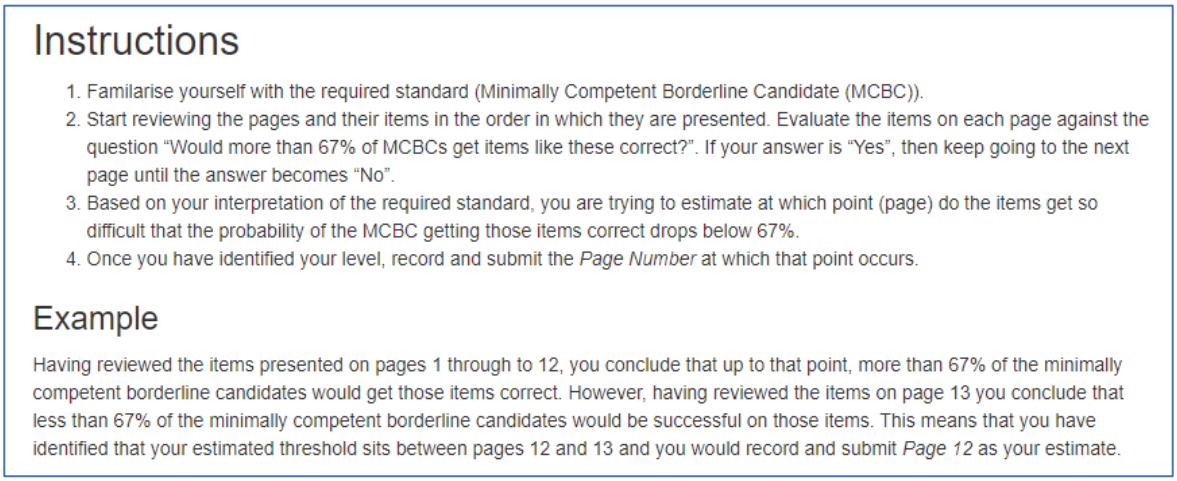
Training was provided to the panel of judges by an experienced medical educationalist with a background in assessment design and delivery for both undergraduate & postgraduate medical assessments. The concept of the minimally competent borderline candidate was explained and discussed during the training. Judges were asked to imagine either a hypothetical borderline trainee, or they know and would consider borderline from their own experience, to improve the judges understanding of the concept. A simulation activity using simulated questions was performed with the judges as part of the training.

Keeping the borderline candidate in mind, the judges were asked to anticipate the performance on each item for the borderline candidate; that is, the probability that the borderline candidate would answer the item correctly. After making their individual initial ratings, judges then shared and discussed their ratings, along with some performance data provided for the items. After some discussion, judges were asked to make a second estimate, allowing them the opportunity to either retain or submit their original estimate, or to amend it. These final ratings were then used to determine a recommended cut-score, as previously described (Chapter 3.1.2).

3.3.4 Standard Setting - Modified Bookmark method

Training was proved to the panel of judges by an experienced educator and psychometrician. The Judges were asked to imagine a hypothetical borderline trainee. The judges' task in the bookmark method was not the same as the judges' task in Angoff. In bookmark, the judges are making a judgement on probability. They were asked to consider the candidates Response Probability (RP), that is the probability that the borderline candidate will answer a test item correctly (Huynh 2000).

As previously described, RP67 was chosen for this evaluation, and this was explained to the judges. The concept of the minimally competent borderline candidate (MCBC) was explained and discussed. The trainer checked that the judges were satisfied that the classification point between competent/not competent exists within the available scores. The trainer asked the judges to consider their acceptable range of (raw) cut-scores for the assessment i.e. the maximum and minimum acceptable raw cut-scores. A group range within which the group believes the cut-score should lie was agreed. The following instructions were given to the judges with the IOB.



Instructions

1. Familarise yourself with the required standard (Minimally Competent Borderline Candidate (MCBC)).
2. Start reviewing the pages and their items in the order in which they are presented. Evaluate the items on each page against the question "Would more than 67% of MCBCs get items like these correct?". If your answer is "Yes", then keep going to the next page until the answer becomes "No".
3. Based on your interpretation of the required standard, you are trying to estimate at which point (page) do the items get so difficult that the probability of the MCBC getting those items correct drops below 67%.
4. Once you have identified your level, record and submit the *Page Number* at which that point occurs.

Example

Having reviewed the items presented on pages 1 through to 12, you conclude that up to that point, more than 67% of the minimally competent borderline candidates would get those items correct. However, having reviewed the items on page 13 you conclude that less than 67% of the minimally competent borderline candidates would be successful on those items. This means that you have identified that your estimated threshold sits between pages 12 and 13 and you would record and submit *Page 12* as your estimate.

Figure 3.4 Instructions given to judges undertaking the Bookmark Method

The judges were given the Ordered Item Booklet and instructions to guide the standard-setting process.

The bookmark placement proceeded to three rounds with discussion in between (Mitzel, Lewis et al. 2001). The judges were provided with the Ordered Item Booklet (OIB). The pages were arranged in order of increasing difficulty/performance. Each page represents a particular candidate ability level.

The items on each page were the items that candidates represented by that ability level have a 67% probability of answering correctly, or, 67% of candidates at that ability level

would answer the items correctly. For each page, the judges considered whether they believed that a single MCBC had a probability of >= 67% of answering the test items on the page correctly. If the judges answer 'yes', they moved to the next page and repeated the process. If the judge answered 'no', they stopped and recorded the page number immediately previous.

At the end of the first round, the bookmarks were placed without discussion between judges. During the second round, judges were provided with information from round 1, including the range of the cut-scores from the different judges, the mean and median cut-score. The new, round 2 cut-score was calculated and the new median for the panel of judges presented. During round three, data around the impact of the cut-score, that is the percentage of candidates that were passing, was presented. Judges were encouraged to discuss what makes each item in the OIB progressively more difficult than the previous item (Mitzel, Lewis et al. 2001). At the end of round 3, the judges made their final decision on placement of the bookmark and the cut-score was calculated by taking the mean of the ability levels for the page numbers submitted by all the judges.

The group then discussed the findings from the Angoff standard setting process, which had been performed by the same group of judges on an earlier date. When the two standard setting methods were complete, the process and the results were evaluated.

3.4 Results

3.4.1 Selection and training of the judges

The panel of judges was comprised of five judges from four countries, Ireland, England, Norway and the Netherlands. Three judges were male and two were female. There was a difference of twenty years in the age range between the youngest and the oldest judge. Broadly speaking, there were some differences in specialist interest areas among the judges when considering the four key pillars of clinical microbiology, laboratory, Infection prevention and control and antimicrobial stewardship. Some members of the group were entirely clinical, and some had academic commitments. One member of the group had a postgraduate Diploma in Medical Education. All members of the group were involved in postgraduate education at a national level in their own country. All members of the group were involved in postgraduate education development in Europe as members of the UEMS section of MM.

3.4.2 Assessment delivery and Psychometrics:

Assessment delivery

The final assessment delivered to the candidates was a Multiple-Choice Question (MCQ) assessment, containing 99 single best answer (SBA) items.

All candidates completed the assessment well within the 150 minutes allotted.

Thirteen items from the test were selected for review by the panel based on the psychometric analysis. Most of these items were selected for review because they were very difficult. Following discussion, four items were suppressed so the final number of items in the test was 95. One item was remediated and was scored; this is discussed in more detail in Figure 3.11.

Candidate demographics

96 candidates plus one scrutineer sat the exam. There was representation from 18 different countries. Fifty-nine candidates were trainees and 36 were Consultants. Two candidates did not specify whether they were a consultant or trainee.

Assessment psychometrics

Cronbach's alpha (Scale reliability coefficient) for the assessment was 0.82. A value of >0.80 is expected for a high-stakes exam (Hecker and Violato 2009). The mean test score was 49.47 (range 28 to 78). The Standard Error of Measurement (SEM) was 4.22. This was acceptable, so the test was deemed valid. There was an optimum distribution of item difficulty and candidate ability leading to a very discriminating assessment (See Figure 3.5).

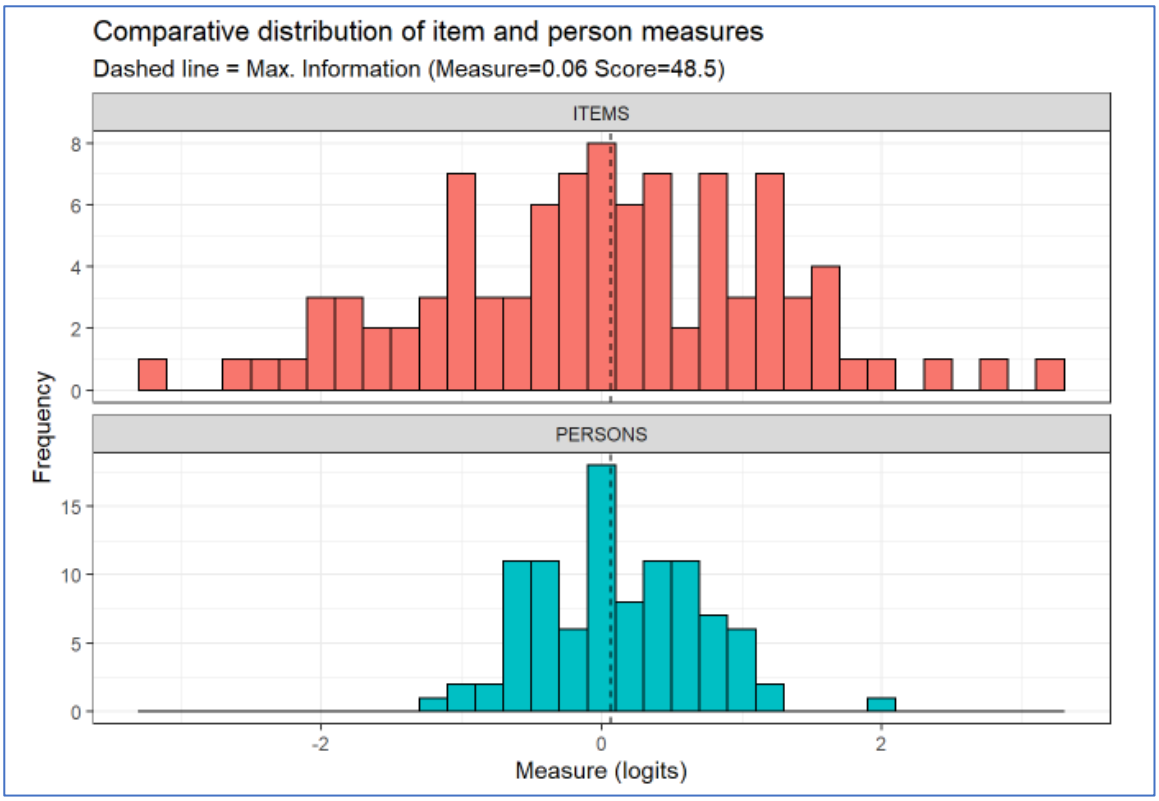


Figure 3.5 Person and Item Ability Distribution

Rasch analysis indicated a good model fit with a person reliability value of 0.82 and an item reliability value of 0.96. There was an optimum distribution of item difficulty and candidate ability leading to a very discriminating assessment.

The test items were placed on an item difficulty map and showed a good mixture of easy and difficult items (Figure 3.6).

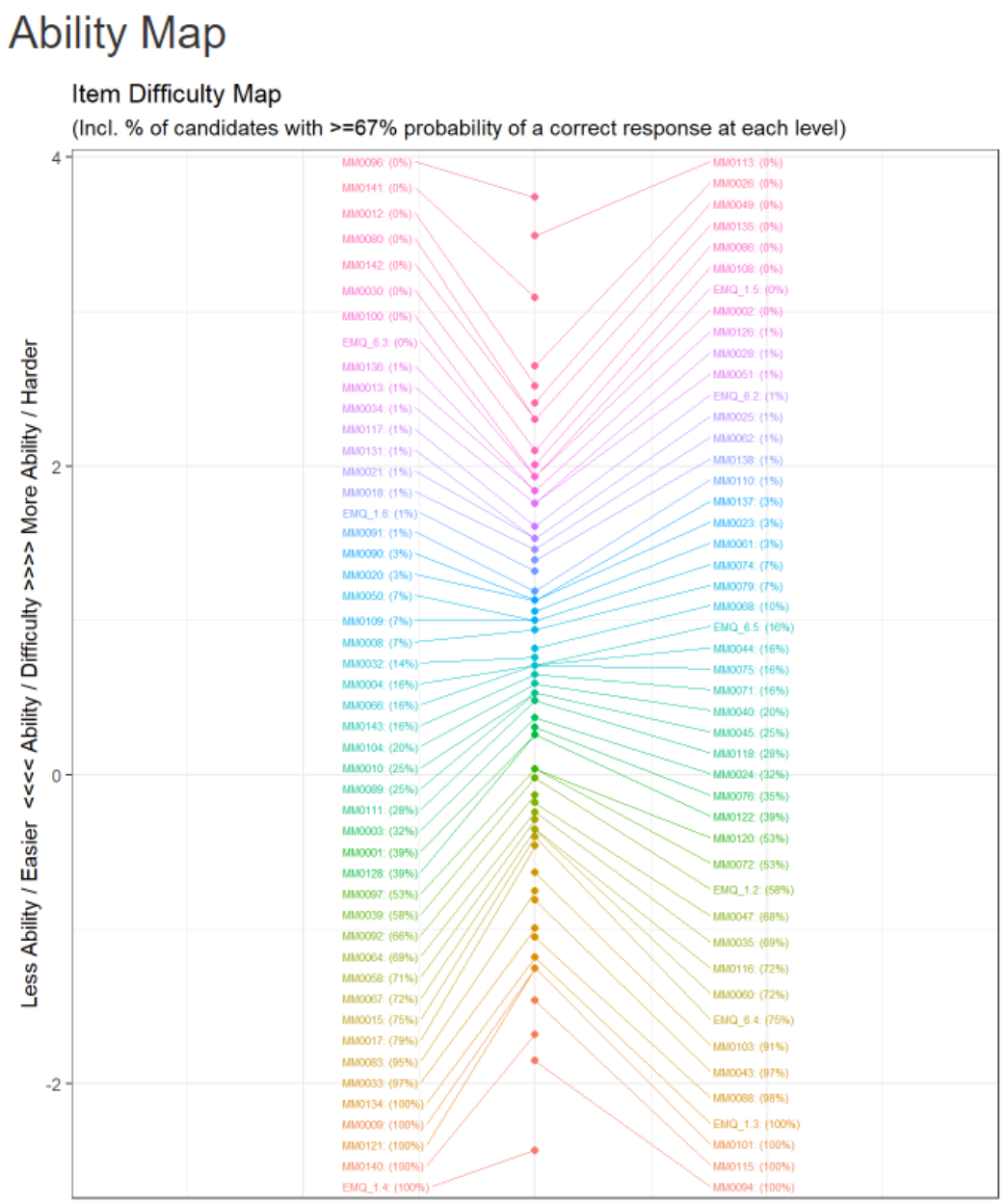


Figure 3.6 Item difficulty map

The questions were mapped according to their difficulty. The item difficulty map shows a good range of questions from easy (bottom of the map) to difficult (top).

Test items were analysed to look for 'Flags' that might indicate test items with a higher probability of a problem.

Item fit discrepancy can be measured using mean-square residual summary statistics, such as OUTFIT and INFIT. These statistics have expectation of 1.0 (perfect fit) and range from 0 to infinity. When the mean-squares is greater than 1.0, this indicates underfit to the Rasch model, i.e., the data are less predictable than the model expects. When the mean-squares less than 1.0, this is an indication of overfit to the Rasch model, i.e., the data are more predictable than the model expects (Wright BD and JM 1994). The Rasch Model Fit Metrics looked for items outside the parameters of 0.7-1.3. These parameters that are deemed reasonable for an MCQ exam (Wright BD and JM 1994). Four items 'flagged' which were subsequently reviewed (Figure 3.7).

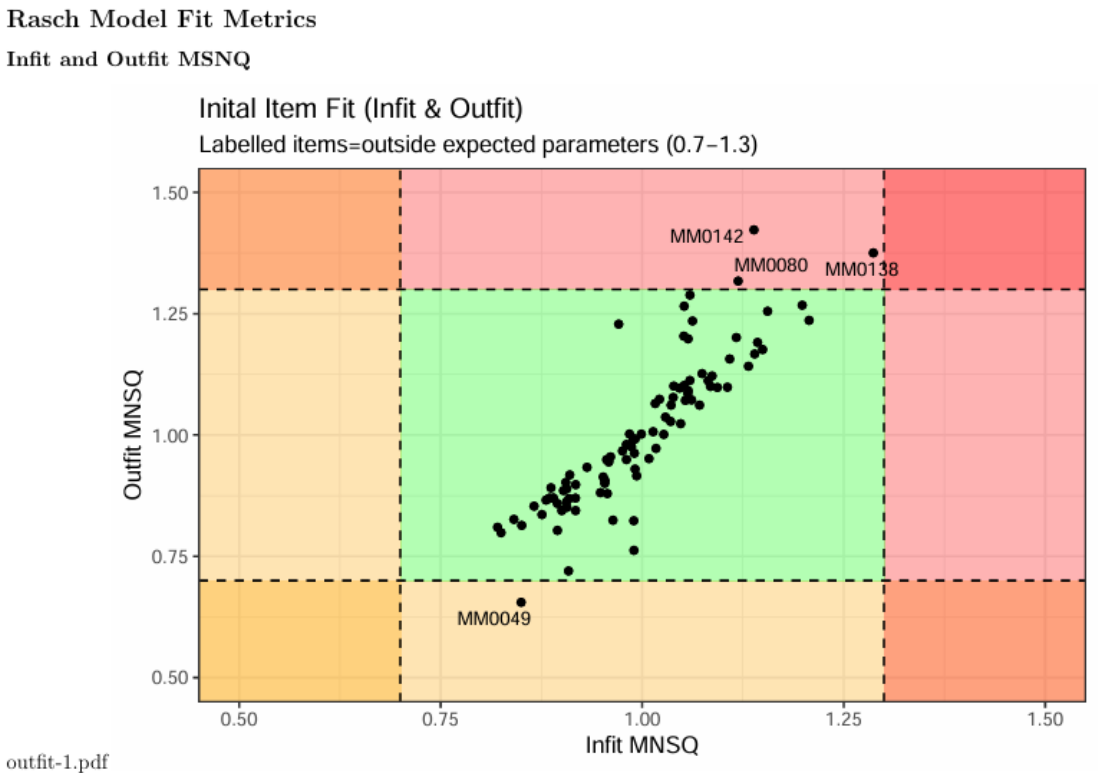


Figure 3.7 The Rach Model Fit Metrics

An analysis of initial item fit flagged 4 items that fell outside the expected parameters of 0.7-1.3.

Item performance statistics for difficulty and discrimination were reviewed looking for 'flags'. Thirteen items were flagged based on the item Performance statistics (Figure 3.8). Most of the items were selected because they were unduly difficult, and all flagged items were reviewed by the panel of judges. Difficulty scores of <0.20 indicates that the test item was difficult (Sabri 2013). The value of the discrimination coefficients ranges between -0.1 and 0.1 (Sabri 2013). The discrimination point biserial value of <= 0.2 indicates items that need review. The items with a negative discrimination need review.

Item Performance Statistics

Suspect Items For Review

Table 1: Item Performance Statistics for items that should be reviewed

Item	Difficulty	Discrimination	Flag
EMQ_6.4	0.04	-0.16	X
MM0084	0.04	0.05	X
MM0096	0.05	-0.12	X
MM0113	0.07	-0.01	X
MM0141	0.10	0.09	X
MM0105	0.14	-0.09	X
EMQ_6.6	0.15	-0.11	X
MM0026	0.15	0.06	X
MM0012	0.18	0.02	X
MM0135	0.20	0.04	X
MM0142	0.21	-0.13	X
MM0080	0.21	-0.07	X
MM0048	0.21	0.05	X

Figure 3.8 Item Performance statistics of items flagged for review

The 13 items 'flagged' for review. The difficulty scores of <0.25 indicates that the test item was difficult. The discrimination point biserial value of <= 0.2 indicates items that need review. The items with a negative discrimination and/or very high difficulty need review.

The distractor analysis 'flagged' seven items that needed a review (Figure 3.9). The distractor analysis considers the mean ability of the candidates. Items will be 'flagged' if the distractors are chosen by the strong candidates.

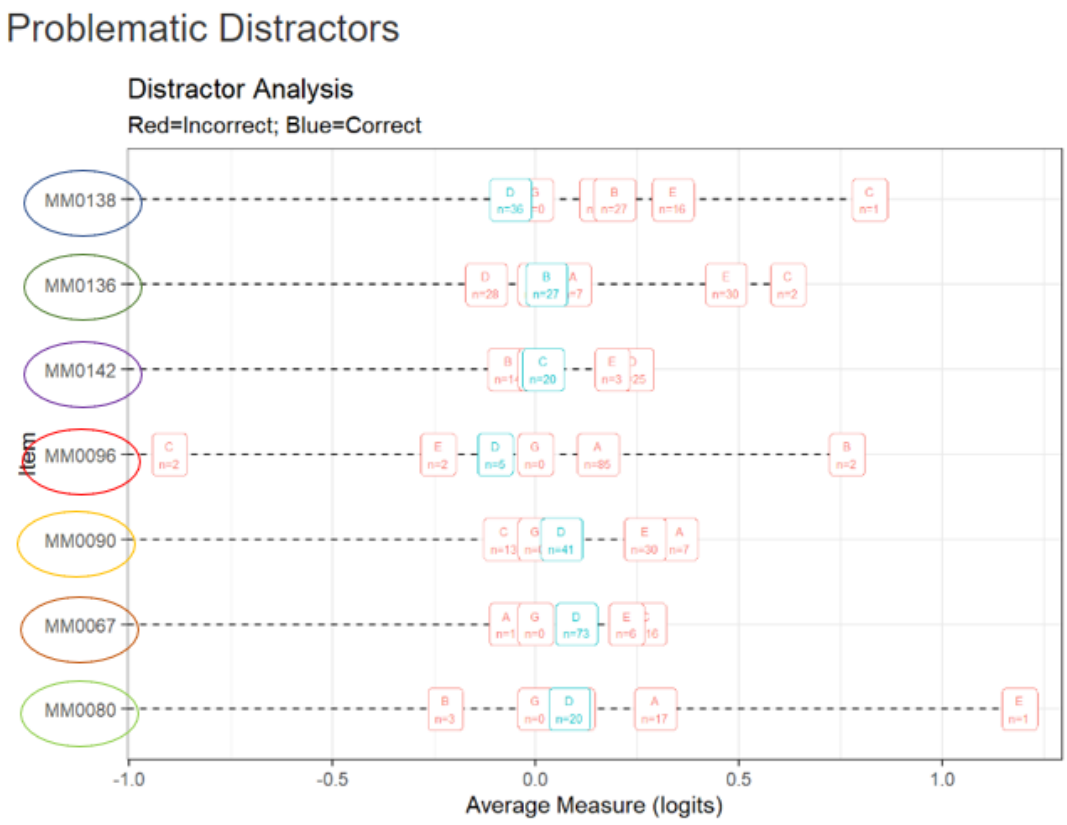


Figure 3.9 Analysis of problematic distractors.

Item statistics for all items were then plotted to facilitate identification of items where either item difficulty/facility or item discrimination (corrected point biserial) fell below recommended thresholds, as depicted in the red areas of the chart in Figure 3.10.

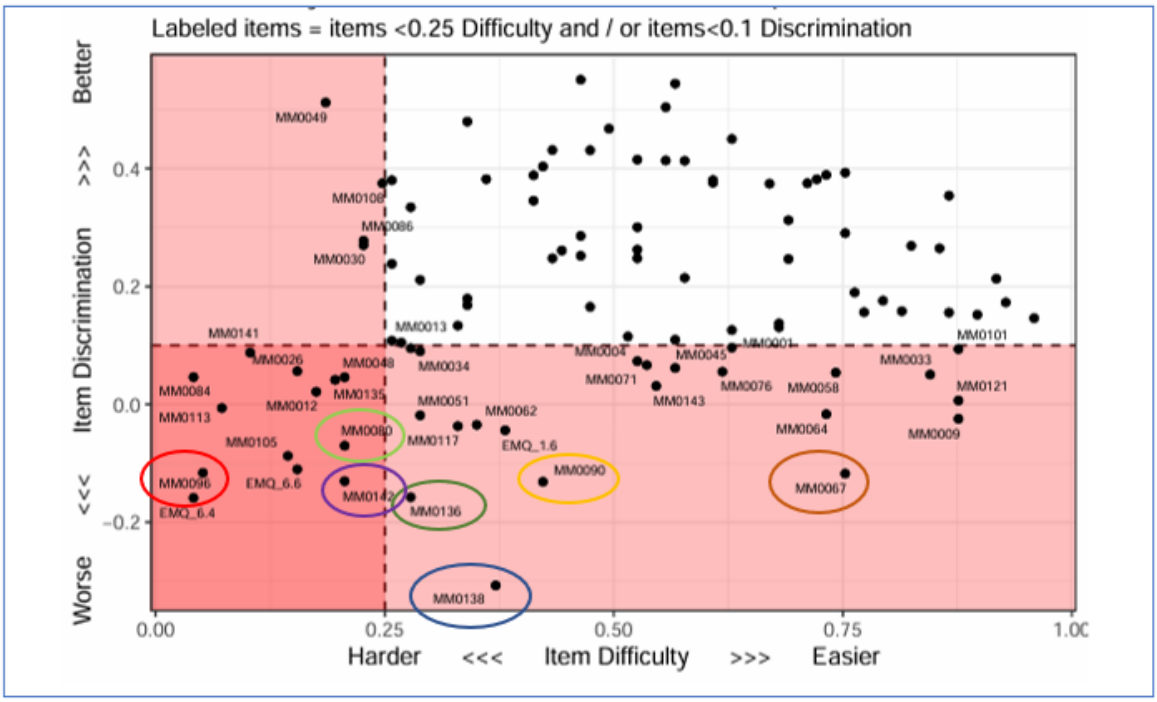


Figure 3.10 Item Difficulty & Discrimination

The plot shows item difficulty plotted against item discrimination (corrected point-biserial). Items which fall below relevant thresholds fall into red areas of the chart and are labelled. 13 items from the red area (bottom, left) were selected for review by the panel of judges. I have also highlighted the seven items 'flagged' based on the distractor analysis (Figure 3.9).

Question	Comments
MM0138	This question has very poor discrimination. This item is also very difficult (low item difficulty)
MM0136.	Poor discrimination and very difficult
MM0142	Very difficult. Discrimination poor.
MM0096.	Extremely difficulty. Discrimination poor.
MM0090.	Poor discrimination, but item difficulty better.
MM0067.	Poor discrimination, higher item difficulty/faculty, so less difficult
MM0080.	Poor discrimination, poor item difficulty.

Table 3.1 Seven problem questions

This table discusses seven problem questions in more detail.

One item 'flagged' on four of five categories in the analysis (Figure 3.11). It also has a very negative discrimination (which means weaker candidates are more likely to get this item correct than stronger candidates). This item also exhibits a great deal of 'noise' which means that the results for this item are very unexpected. The correct option for this item is listed as Option E but few candidates (n=4) seem to have chosen this option and those that have tend to be weaker overall. The candidates with the highest overall ability seem to have chosen Option C. Results like this might suggest that this item might have been mis-keyed. The item was reviewed, and it was agreed that option C should be the correct option rather than option E. The key should be corrected, and the items re-scored.

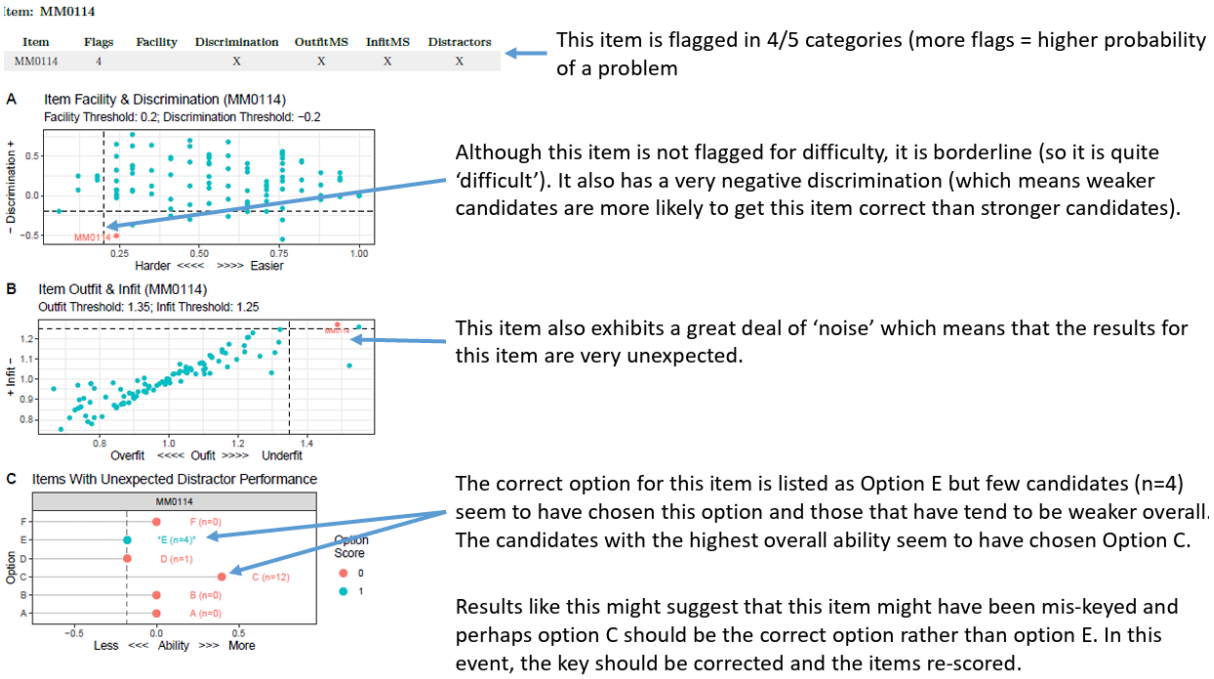


Figure 3.11 Exploration of item MM0114.

Item MM0114 flagged in 4 of 5 categories assessed; discrimination, outfit, Infit and distractors and was borderline for difficulty. Analysis suggested that the item was 'mis-keyed'. Option C should be the correct option rather than option E. The key was corrected, and the items re-scored.

The candidate responses were analysed for non-standard responses. Five candidates were identified with >10% non-standard responses (Figure 3.12). The red squares indicate when a question was skipped. It is unusual to see candidate skip a question when the test is not negatively marked. If there were a lot of candidates with skipped questions towards the end of the test e.g. Candidate D, this can indicate that the test is too long, and candidates do not have enough time to finish.

The candidate feedback survey (this will be discussed in Chapter4) indicated that when candidates completed the last question, the test finished. This meant that if a candidate

skipped and question and planned to go back to it later, they could not once they completed the last question. This occurred without warning the candidates that it would occur. Candidates skipped questions and planned to go back to them at the end, which would explain the pattern observed. This was corrected in the next exam.

Candidates with >10% non standard responses

There were 5 candidates with >10% non-standard responses. Non standard responses are unexpected characters or blank/ missing responses.

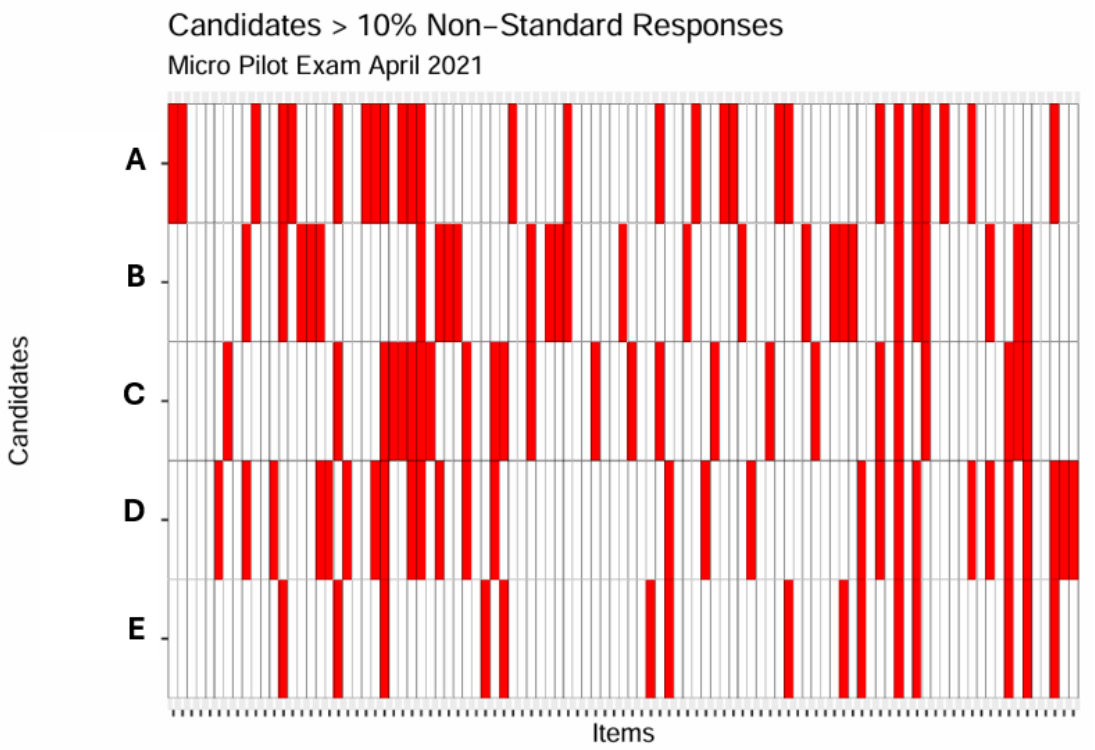


Figure 3.12 Candidates with >10% non-standard responses

Five candidates A-E were 'flagged' due to non-standard responses.

3.4.3 Standard Setting - Angoff

Using the Angoff estimates provided, the cut-score was calculated as **55.115** out of a total of 95 (58.02%; Figure 3.13). When applied to the raw data, the pass rate would have been **28.87%** (Table 3.2; Figure 3.13)**.** The correlation between the Angoff estimates and item facility was weak (r^2 = 0.24; Figure 3.14). Applying the Angoff estimates, the pass rate would have been 28.87%.

Result	Number of candidates (n = 96)	Percent
Fail	69	71.13
Pass	28	28.87

Table 3.2 Summary of overall results with Angoff cut-score

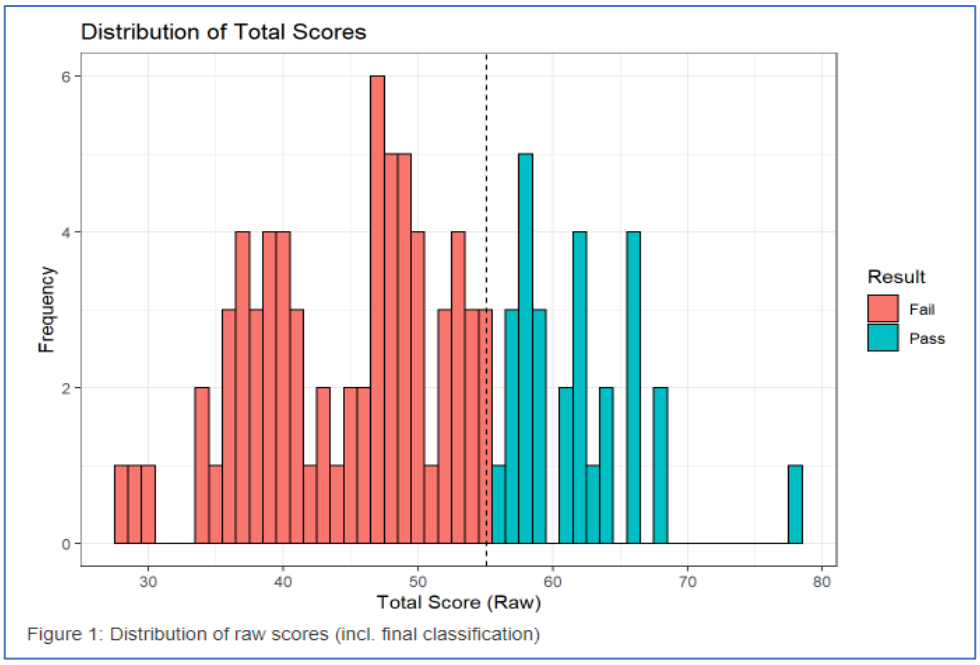


Figure 3.13 Summary of overall results with Angoff cut-score

This figure shows the distribution of raw scores using the Angoff Method to calculate the cut-score. The cut-score was calculated as **55.115** out of a total of 95, a passing score of 58.02%. The scores of candidates that did not achieve the passing standard are in red.

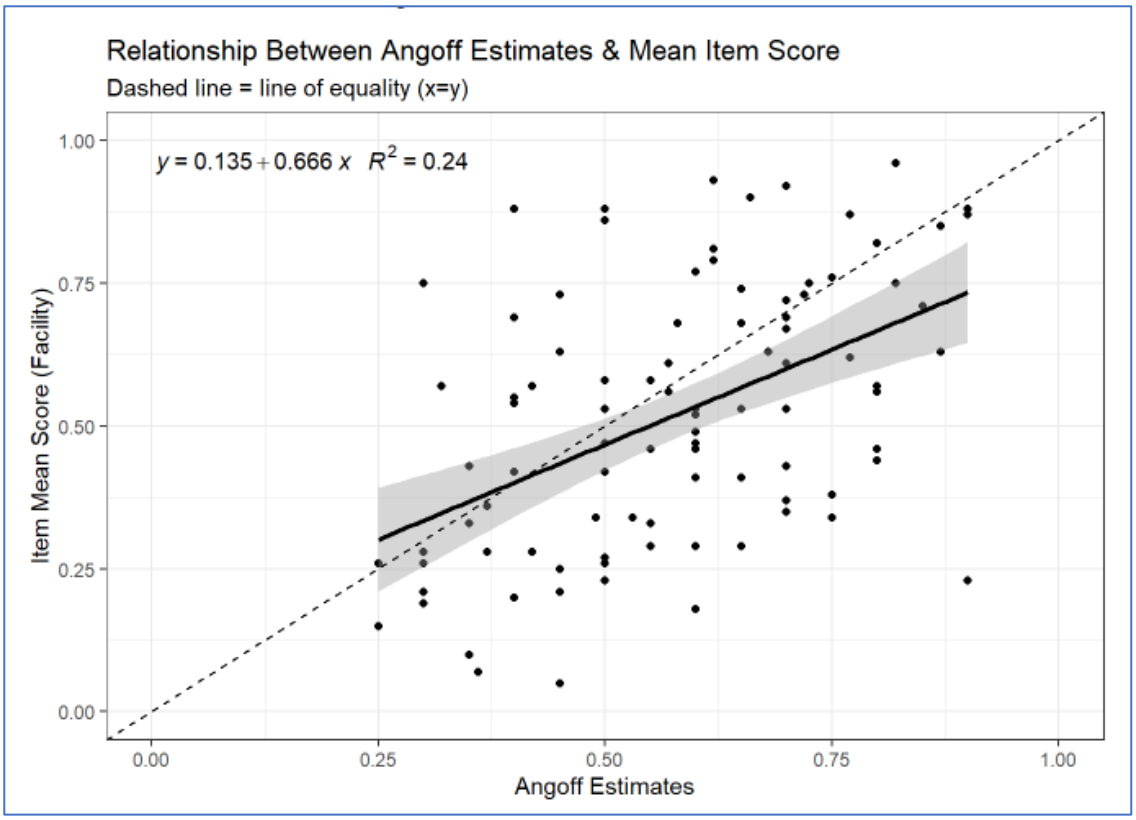


Figure 3.14 Relationship between Angoff estimates and Item Facility.

The Angoff estimates were compared to the Mean Item Scores for each test item. The relationship between Angoff estimates and Item Facility was weak ($r^2 = 0.24$).

The Rasch probabilities were compared to the Angoff estimate. As expected, the raters (Angoff method) overestimated the difficulty of easy items and underestimated the difficulty of the difficult items (Impara and Plake 1998) (Figure 3.15).

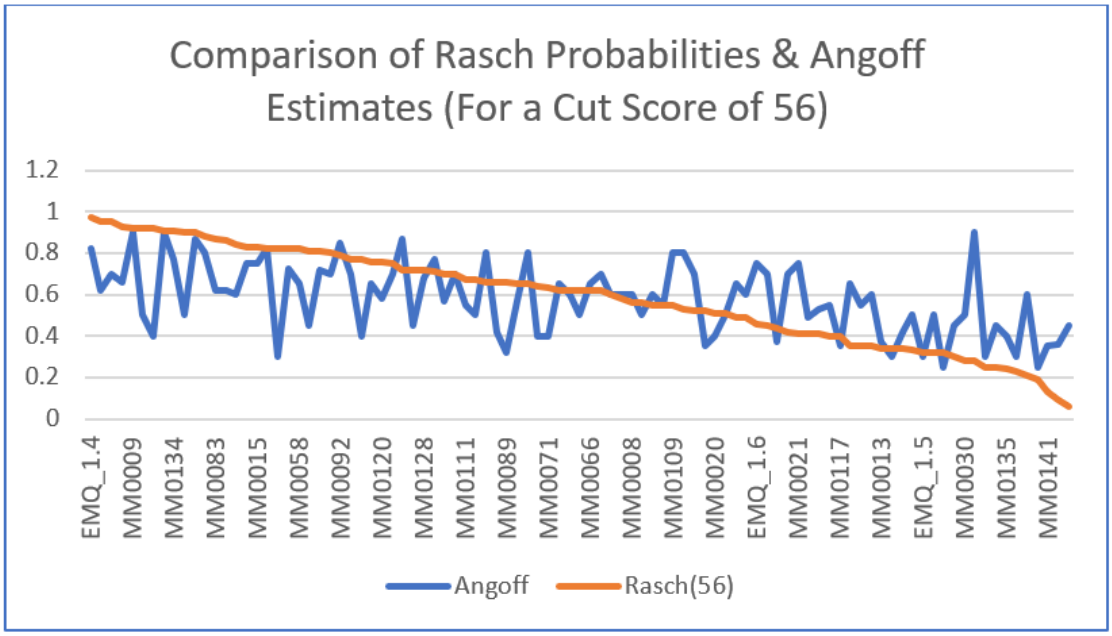


Figure 3.15 Comparison of Rasch Probabilities and Angoff Estimates (For a Cut-Score of 56)

The Rasch Probabilities derived using IRT were compared to the Angoff estimates as assessed by the judges.

3.4.4 Standard Setting - Modified Bookmark method

Using the Bookmark method, the cut-score was 48 out of 95 (50.53%; Figure 3.16). When the bookmark cut-score was applied to the candidates results, 68% of candidates passed (Figure 3.17).

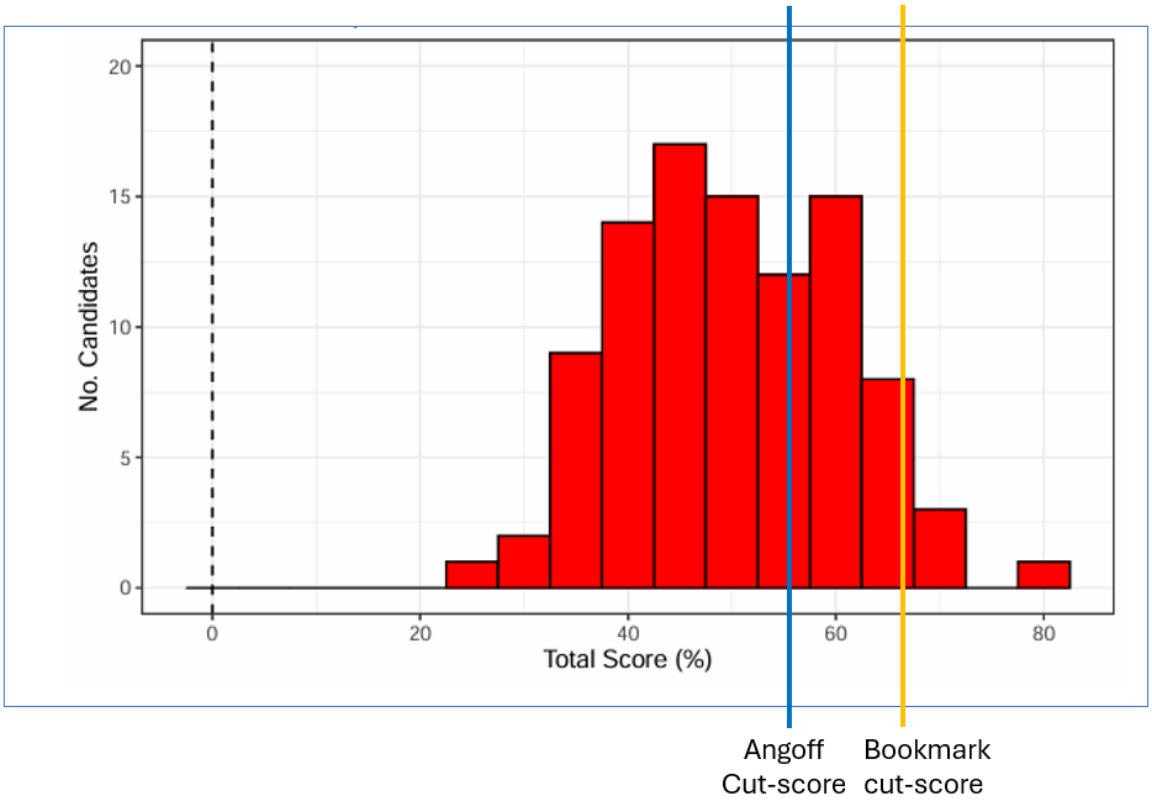


Figure 3.16 Cut-scores for Angoff and Bookmark Methods of Standard Setting

Using the Angoff estimates, the cut-score was calculated as 55.115 out of a total of 95. Using Bookmark, the cut-score was 48 out of 95.

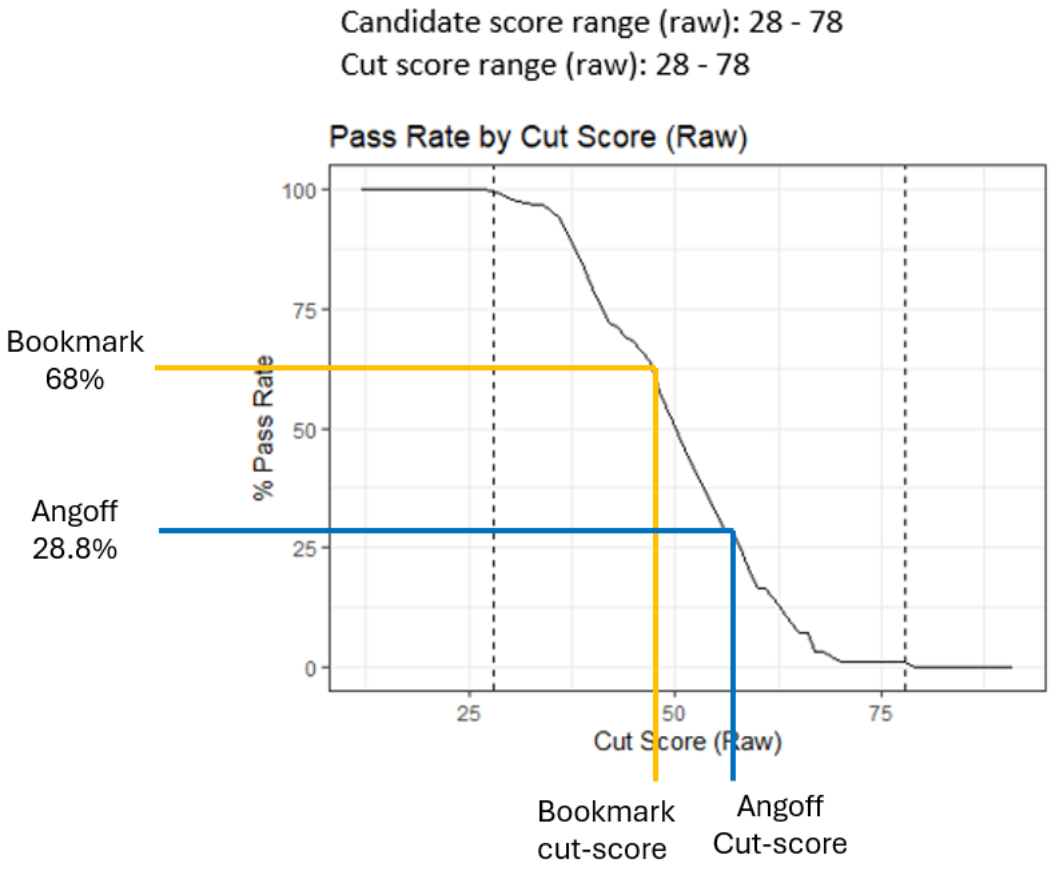


Figure 3.17 The Passing rates for Angoff and Bookmark Standard Setting Methods

The Angoff cut-score was 55.11, resulting in a passing rate of 28.8%. The Bookmark cut-score was 48, resulting in a passing rate of 68%.

3.4.5 Reporting of results & provision of feedback to candidates

The results were further analysed to see how candidates performed in the 9 sections of the curriculum. The dashed line indicates the required, passing standard determined by bookmark (Figure 3.18).

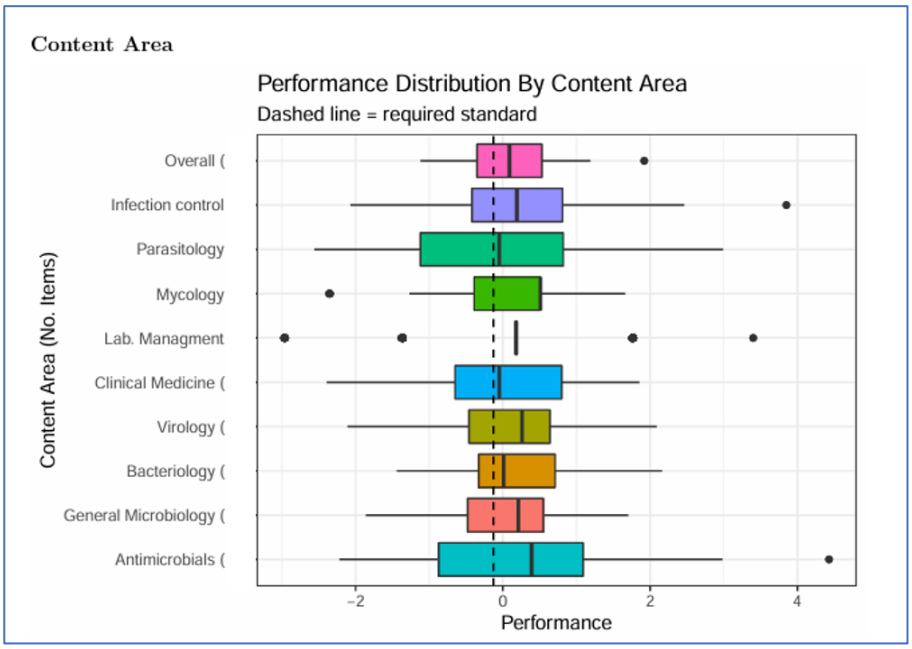


Figure 3.18 Performance Distribution by Content Area

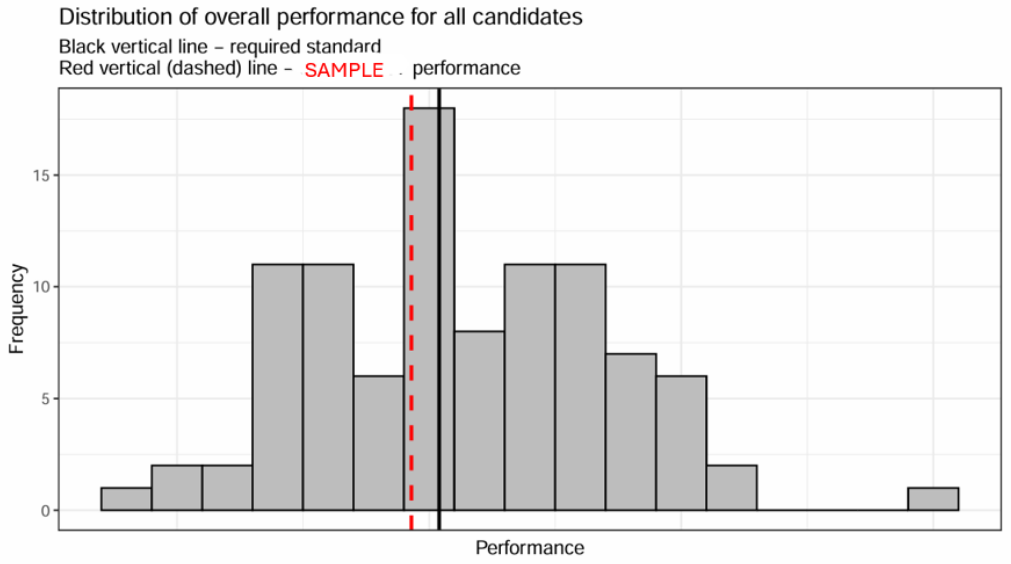
The figure shows overall performance by candidates (pink) and the performance by candidates in the nine areas of the curriculum. The passing standard is indicated by the dashed line.

A feedback report was generated for each of the candidates sitting the pilot. The report indicated their overall performance, whether they met the required (passing) standard overall and whether they met the required standard in each of the 9 curriculum areas. The provided examples (Figure 3.19) are from a sample feedback report for a candidate that did not achieve the performance standard required to pass the test. The candidate's performance in each area of the curriculum is summarised.

Overall result: Standard Not Met

You achieved a raw score of 46 out of a possible maximum score of 95. The raw score that represents the required standard on this assessment is 48. Note that the required standard remains constant over different diets of the assessment but the raw scores that represent the standard may vary due to the normal variation in overall difficulty of different diets.

Overall



By Content Area

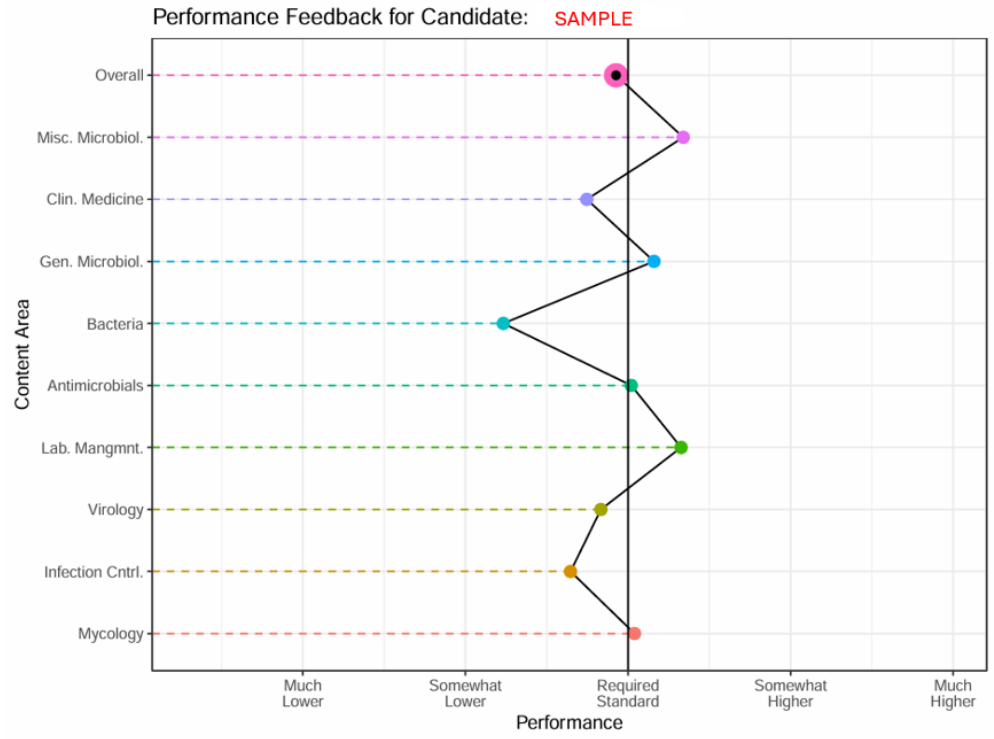


Figure 3.19 Feedback report for Candidate

The feedback report for candidates had two parts. The first indicated whether the standard was met and shows the candidates overall performance, compared to the other candidates. The second part of the report shows the candidates' performance in each area of the curriculum. In this example, the candidate did not meet the standard required to pass. The candidate's weakest areas were bacteriology and infection control.

The candidate feedback report should be useful for candidates wishing to retake the exam, by helping them to prioritise areas to study.

3.5 Discussion

> *"There is no perfect 'cut-score' and only by chance is any the test score the "true" score for a test-taker*" (Zumbo 2016).

It is more realistic, therefore, to view the process as one that is based on probability and demonstrate the validation steps and processes that were undertaken to arrive at the score, i.e. the evidence trail to support that it is not an unreasonable score (Zumbo 2016).

In this education simulation, we chose to use two standard setting methods, Angoff and Bookmark, both of which are well-established, criterion-referenced, test-centred methods with an extensive evidence-base in support of their use for our test format of MCQs (Angoff 1971, Livingston and Zieky 1982, D.M., Mitzel et al. 1996, Bandaranayake 2008). The use of an *absolute standard* reflects a desired level of mastery and therefore is recommended for a high-stakes exam (Downing and Yudkowsky 2009). The specified process for each standard setting method was followed, to compare the experience and outcomes and agree a method for the European Exam in MM that is "f*easible, academically and legally defensible and otherwise fit for purpose*"(Coombes, Roberts et al. 2016).

Psychometric analysis plays an important role in quality assuring that the interpretation of the assessment outcomes are reasonable and is an important source of validity evidence (Coombes, Roberts et al. 2016). For our assessment, Rasch analysis indicated a good model fit with a person reliability value of 0.82 and an item reliability value of 0.96. Cronbach's alpha was 0.82, demonstrating that the examination had very good internal consistency. For high-stakes tests, reliability co-efficients are generally expected to exceed r_{xx}=0.80. This means that consistent and stable measurements are responsible for 80% of the variance

in repeated measurements (Hecker and Violato 2009). The Standard Error of Measurement was 4.2. An SEM of =< 5.0 is usually considered acceptable.

The psychometric analysis in bookmark identified 13 items for review (Figure 3.8 and Figure 3.10). Following review by the judges, 8 items were deemed acceptable, but difficult. One item had an error (Figure 3.11); the item had been 'mis-keyed', this was corrected and could still be scored. Four items were removed because the judges concluded that the terminology used might have inconsistent meanings in different countries. This was very useful information and will be used to guide the writing in the first diet (sitting) of the real exam. The final number of items scored was 95.

Analysis of the candidate response for candidates >10% non-standard responses identified 5 candidates for review due to skipped questions (Figure 3.12). Using the feedback from the candidate questionnaire, we concluded that this occurred because when candidates completed the last question, the test finished. This occurred without warning the candidates that it would occur. The likely explanation is that candidates skipped questions and planned to go back to them at the end. This was issue was addressed in the next exam.

> "*The key to defensible standard-setting lies in the choice of credible judges and in the use of a systematic approach to collecting their judgement*" (Downing and Yudkowsky 2009).

The same judges were used for both standard setting methods. The credibility of the assessment will be determined by the number and nature of the judges (Norcini 2003). An essential component to any standard setting method is the selection of the judges (Norcini 2003) and ensuring that these judges then receive appropriate training and information in

preparation for the standard setting process, to ensure consistency and reliability. The agreement among the judges is important for the reliability of the performance standard (Friedman Ben-David 2000). Selection and recruitment of sufficient numbers of qualified judges is a challenge for any process of this nature. There were five judges participating in our standard setting, which is deemed acceptable (if low); Norcini and Shea recommend five to ten judges (Norcini and Shea 1997), while others would recommend a minimum of five to thirty (Zieky and Livingston 1977). Other authors prefer larger groups, and have concluded that ten to fifteen is optimal (Hurtz and Hertz 1999).

Jaeger describes the qualifications that the judges should have i.e. that they should be experts in the field related to the examination, familiar with the exam methodology, good problem-solvers, familiar with the level of the exam candidates and have an interest in education. They should also have the ability to conceptualize (Jaeger 1991). It has been shown that the mix of judges in the committee may be more important than the actual number, and there should be a good mix in terms of gender, race, age and professional roles (Norcini 2003, Hambleton and Pitoniak 2006). Although we had only five judges in our group, we ensured that we had a good mix of demographics, specialty interests and experience. All our judges were content experts and experienced postgraduate trainers who know and understand the target population well, characteristics which Downing believes are the most important in a judge (Downing and Yudkowsky 2009).

The judges' clinical and academic experience may affect the cut-score. The Angoff method of standard setting has been linked to resulting in a low cut-score and consequently a high passing rate (Wayne, Fudala et al. 2005), but the pass mark may increase with an increase in experience and academic degrees among the raters (Rezigalla 2015), so we ensured that our committee of judges included raters with different levels of experience (Rezigalla

2015). Additionally, less experienced judges may tend to set a higher cut-score if they have difficulty in conceptualising the "Borderline Candidate" and revert to a baseline of conceptualising what a candidate "should" achieve if a safe (but low) pass (as would often be the case when examining in assessments), as opposed to what a Borderline Candidate "would" likely achieve.

It is known that judges are likely to reply on opinions of other judges to make their decisions (Dawber 2002), the concern with the judges engaging in this discussion is that one rater may dominate in the group (Hurtz and Auerbach 2003). However, our experience was that the discussion of the judges' ratings with each other, was useful to reach consensus, and also to provide an explanation and promote further thinking with respect to the rating and the borderline candidate ability. This is particularly important in an international examination, where candidates come from different countries, with different scope of practice and perhaps different emphasis on areas of the curriculum in their training programmes.

When the Angoff method of standard setting was applied, the cut-score was calculated as 55.115 out of 95 (58.02%). When applied to the raw data, the pass rate would have been 28.87%. In our examination simulation, the relationship between Angoff estimates and facility was weaker than expected. There is some debate around whether the judges can learn the Angoff method effectively in a single training session and it has been shown that judges with prior experience in the Angoff method demonstrate better consensus in their marking (Boursicot 2006). This may have been the case with our group, all of whom were new to the standard-setting process. The judges might have benefited from more detailed discussion around Performance Level Descriptors (PLD) (Egan, Ferrara et al. 2009) which has been shown to improve the reliability of the process (Yim and Shin 2020).

Many authors have criticised Angoff for producing high cut-scores and attribute this to the unrealistic expectations by judges when they rate item by item (Impara and Plake 1998). There have been several studies comparing Angoff with other standard-setting methods which demonstrated that Angoff produced higher cut-scores. One group compared Angoff method with the borderline method found that Angoff gave more demanding or higher cut-scores (Näsström and Nyström 2008). In this paper, there were multiple cut-scores, for three different grades or categories; fail, pass, and pass with distinction. Comparing the results from the two standard setting methods, 13% of the categorisations differed, e.g. 24% of students categorised as Pass by Angoff method, got a higher grade by Borderline procedure (Näsström and Nyström 2008). This was also the finding in a study comparing the Angoff method to the 'norm-referenced' method. The Angoff method resulted in a higher cut-score, resulting in a marked difference in the pass rate, which was 39% with Angoff and was 88% using the norm-referenced method (Elfaki and Salih 2015).

Another criticism of the Angoff method is that it displays low inter-rater reliability. In a study by Mills comparing Angoff with two other methods, contrasting group and borderline, they found there was a wide range in the judges standards on every test, using Angoff (Mills 1983). Using frame-of-reference rater training strategies may result in significantly higher levels of interrater reliability (Fehrmann, Woehr et al. 1991).

The bookmark method of standard setting, the cut-score was calculated as 49.47 out of 95 (52.07%). When applied to the raw data, the pass rate was 68%. Confidence in the passing score and comfort with the process between the judges in our group varied. Most of our judges employed a strategy, previously described, of scanning a range of items before deciding on the position of their bookmark (Dawber 2002). As a result, some judges had concerns regarding the order of items in the OIB. While it is important to reiterate that the

order is based on the examinees scores in the test they are reviewing, there are some things which may affect the order, for example, if one or two weaker candidates had a 'lucky guess' on a difficult item, this may change the items position in the IOB (Plake 2005). A criticism of bookmark is the requirement for an expert in psychometrics, which may be beyond the resources for some exam groups, however, as demonstrated in the results section, above, the data generated with IRT is very *rich* and has many advantages, including making it possible to bank item statistics for future use.

In our study, the two criterion-referenced (test-centred) methods applied and then compared were Angoff and Bookmark; the cut-score for Angoff was higher than the cut-score for bookmark, resulting in a large difference in the passing rate (Figure 3.17), a finding consistent with many other studies in published literature. One study comparing a Rasch IRT model of standard setting to Angoff, not only found that Angoff consistently had higher cut-scores than the Rasch IRT model, but also reported that Angoff had lower inter-judge consistency and less rater agreement than the IRT model (Wang 2003).

Buckendahl *et al* compared Angoff to Bookmark, and overall, the two methods did achieve similar cut-scores, but with modifications of both methods (Buckendahl, Smith et al. 2002). In Bookmark, they used Classical Test Theory, rather that Item Response theory. They used a variation of Angoff where the judges decide whether a borderline candidate will get the item "right" or "wrong" (Buckendahl, Smith et al. 2002). Bookmark demonstrated a decrease in the standard deviation between round 1 and round 2, this would indicate a higher level of inter-rater judgement. This finding would likely be reassuring for some stakeholders. However, with each round of modified Angoff, the cut-scores dropped, and the standard deviations increased. In Buckendahl *et al*, when the Angoff method was applied, there was no discussion between rounds, but impact data was provided. The

judges described similar levels of confidence and comfort among the judges for the process. (Buckendahl, Smith et al. 2002). Another study comparing Angoff with Bookmark found no difference in cut-scores when the RP67 was used for Bookmark, but then did find a difference when RP50 was applied for Bookmark instead (çetin and Gelbal 2013). Likewise, Yousefi Afrashteh *et al* also demonstrated a lower cut-score for Angoff than with Bookmark when applying RP50 (Yousefi Afrashteh 2021).

It is also important to consider whether the Angoff cut-score is too high, or is the bookmark cut-score too low? When Reckase used simulated data to compare Bookmark to Angoff, a negative bias was reported in the bookmark method, the cut-score for bookmark was lower than the cut-score hypothesised by experts (Reckase 2006). Lewis suggested at a symposium in 2005 (National Council on Measurement in Education, Montreal, Canada, as cited by Karantonis et al, 2006) that a modification of calculating the cut-score in bookmark which may address the low cut-scores reported. It was suggested that the mean of the items above and below the bookmark could be used, rather than that of the item below (Karantonis and Sireci 2006).

To help make the final decision on the standard-setting method, an opinion was sought, by personal communication from Prof Cees P. M. van Der Vleuten, Maastricht University, the Netherlands. The decision was made to apply the cut-score arrived at by the Bookmark method because it considers the test difficulty. For most judges, the grounding in IRT gave them more confidence in the reliability and reproducibility of the results. The increasing use of bookmark over time indicates that users find it practical and feel confidence in its logic and grounding in IRT, giving the judges confidence in the process (Buckendahl, Smith et al. 2002).

The test psychometrics are just one part of the evidence required for a valid and credible assessment. Downing places emphasis on the process followed during standard setting, There are four principles that underpin the process, it should be systematic, reproducible, absolute and unbiased (Downing 2003). We had good process validity, there was training by an expert prior to commencing which has been shown to be an important step (Yim and Shin 2020). We also had a group discussion around the concept of the borderline candidate for both test methods (Yim and Shin 2020). For each of our standard setting methods our expert was different, one was an experienced educator, with experience in assessment methods, the other was an experienced educator and psychometrician. Although procedural validity is important, it does not provide sufficient validity evidence for the standard setting procedure (Kane 1994). Kane suggested that the test method be assessed for external validity (Kane 1994). One method proposed was to compare the results to the standard that the candidates' teachers expected of them. This has been used in a number of papers looking at standard setting. In a European exam, this is not an option, as most of the candidates were unknown to the judges. The other method proposed to assess external validity was to use a second standard setting method.

Other authors have proposed other criteria to assess the validity and defensibility of the cut-scores. Egan describes 6 criteria which can be used to assess bookmark (Egan 2001). These are (1) precedence, (2) sound research base, (3) scrutiny in the academic community, (4) participant training, (5) documentation and (6) joint effort between sponsoring agency and contractor. She concluded that the bookmark method met the criteria when executed appropriately and produces a standard that is valid and legally defensible (Egan 2001).

Another consideration is *consequential validity*, a validity construct, which refers to the effect the assessment will have on stakeholders, this will include positive consequences the assessment may have for trainees in Clinical Microbiology, who may experience different educational benefits as a result of the development of the assessment. Our examinees received detailed feedback on their performance in the test. The feedback report (Figure 3.19) gave the candidate their overall score and provided a breakdown of their standard in each section of the curriculum, which sections they passed and which sections they did not pass. Norcini in his description of the seven criteria for good assessment, includes that the assessment should have an catalytic effect (Norcini, Anderson et al. 2011). Assessment drives the direction of learning. There is evidence to show that *'whatever kind of test matters in the system has a heavy influence on classroom practice*' (Simmons and Resnick 1993). By giving the examinees detailed feedback on the areas of the curriculum that they performed poorly in, the candidate should be motivated to study those curriculum areas, supporting Norcini's recommendation that examinees should receive feedback that fosters ongoing learning and therefore have an educational effect (Norcini, Anderson et al. 2011). The catalytic effect of assessment is very important for unsuccessful candidates, because it supports remediation (Norcini, Anderson et al. 2011). Trainees will focus their efforts on what assessment tasks tell them is important, the educational effect may be that the EEMM will drive harmonisation of training and practice in Europe.

The final step in the standard setting process occurs after the test, it is deciding what to do afterwards (Norcini 2003). Following our experience of the two standard setting methods, in the pilot examination it was agreed that we would adopt the Bookmark method for the first 'real' European Examination in MM. In any high-stakes examination,

to gain credibility, it is important to consider the stakeholders which are; the examinees, the teachers/trainers, the patients, the healthcare system and regulators. The stakeholders expect the assessment to include quality test material, designed by content experts with a systematic standard-setting process and delivered in a secure manner, protecting the content (Norcini, Anderson et al. 2011). We presented our experience of the pilot exam to some of our stakeholders, at the ECCMID conference, a meeting of CESMA and a meeting of the UEMS section of MM. There is growing concern in society around assessment processes, standards and quality assurance in high-stakes assessments in healthcare. Validity in assessment may be regarded as a social imperative, highlighting the need to establish the quality of newer assessment processes and approaches (Marceau, Gallagher et al. 2018).

As Kane and Downing suggests, there is no on perfect cut-score (Kane 1994) and no 'gold standard' for the process (Downing, Tekian et al. 2006), but I believe we achieved what Zumbo describes as a more realistic approach or expectation, an evidence-based approach,

> "*tied to test validity and establishing an evidential trail that supports that the proposed performance standards and cut-scores are not unreasonable*." (Zumbo 2016).

Chapter 4: Feasibility, evaluation, outcomes

4.1 Background

Pilot study - a study of feasibility

The Concise Oxford Thesaurus defines *a pilot project or study* as an *experimental, exploratory test, preliminary, trial or try out* investigation, cited from Thabane *et al* (Thabane, Ma et al. 2010). While many researchers may use the term 'pilot study', to describe e.g. a small single-centre study, and quite often the emphasis is on the statistics, feasibility is the main focus or emphasis of a pilot study (Thabane, Ma et al. 2010). Other authors exploring the issue of terminology concluded that "*All pilot studies are feasibility studies, but not all feasibility studies are pilot studies.*" (Donald 2018). In the context of social science research, it is recognised that the pilot study can give advance warning of where a project might fail or whether the proposed methods are appropriate and to identify potential practical problems (van Teijlingen and Hundley 2002). All of these are relevant in the context of developing a new high-stakes assessment.

Thabane *et al* believe that many pilot projects are poorly designed, with no clear feasibility objectives, no clear plan for analysis and no clear criteria for determining the success of feasibility (Thabane, Ma et al. 2010).

In the context of Clinical and Translational Research, Moore *et al* define a pilot study as a " *as preparatory studies designed to test the performance characteristics and capabilities of study designs, measures, procedures, recruitment criteria, and operational strategies that are under consideration for use in a subsequent, often larger, study*" (Moore, Carter et al. 2011). This can be applied to a pilot exam, the purpose of which is to also to assess

the designs, measures, procedures, recruitment criteria, and operational strategies that are under consideration for use in the subsequent, high-stakes exam. Moore *et al* make recommendations for investigators considering a pilot study, which include carefully specified aims and objectives, justification of sample size and consideration of how the pilot will relate to the future study or plan (Moore, Carter et al. 2011).

The reasons for undertaking a pilot or feasibility study have been summarised by Van Teijlingen *et al* (van Teijlingen and Hundley 2002). The reasons are grouped under four broad headings, the process, the resources, management and scientific. The four broad headings can be applied to explore the feasibility of setting up a European examination, through a pilot exam. The pilot exam looked at the processes necessary to deliver a European Exam in Clinical Microbiology. It studied all steps in the *process*, from candidate communication, application, registration, blue-printing, question-writing and selection, exam delivery, standard setting and communication of the outcome to candidates. The pilot examined the *resources* that would be required, human and financial, to deliver a high-stakes pan-European Examination and whether that was achievable by the small UEMS section of MM. The pilot explored *management*, specifically the management of the questions, creating a secure question bank and the secure management of candidate data. The fourth heading is *scientific*. In the context of the pilot examination, the processes were guided by the international literature descriptions of best practice approaches to quality in assessment and in particular, the exploration of more than one standard setting method, during the pilot examination. This is discussed in Chapter 3 of this thesis.

Shanyinde *et al* undertook a literature review of published pilot/feasibility studies and found that the majority did not primarily address methodological issues and in that in most papers while discussion around planning did include lessons learned, they judged the

extent of this discussion as minimal (Shanyinde, Pickering et al. 2011). They have created a checklist of methodological issues for feasibility research. While the checklist was developed to look at pilot and feasibility in the context of randomised controlled trials (Shanyinde, Pickering et al. 2011), half of the checklist items could also legitimately apply to assessing feasibility in the context of a pilot examination.

Feasibility study criteria have been used previously to assess other medical education interventions. Jujo *et al* describes a pilot study to assess curriculum feasibility for the development of a medical student cardiac Point-of-care ultrasound (POCUS) curriculum (Jujo, Lee-Jayaram et al. 2021). The authors considered three key messages regarding feasibility; (1) What uncertainties existed regarding the feasibility? (2) What are the key feasibility findings and (3) What are the implications of the feasibility findings for the design of the main study?

Following the feasibility study, significant problems may be identified that may have implications for the project. Bugge *et al* have developed a process for making decisions about the future study, based on the findings of the pilot and feasibility study. They have developed a methodological tool, "A process for Decision-making after pilot and feasibility trials" (ADePT) (Bugge, Williams et al. 2013). The process described is a systematic identification and appraisal of problems that became apparent during the feasibility study and identification of potential solutions. The three-step process starts by assessing whether the problem is likely to affect only the pilot study, only the real study or both. In the next step a range of solutions and the evidence to support them are considered. In the final step, an assessment is made to determine the best option and whether this option is likely to be effective outside the pilot study. This systematic approach to problem-solving

could work equally well when assessing feasibility in a pilot exam, improving the transparency and rigor of stakeholder decision makers (Bugge, Williams et al. 2013).

The criteria used to assess a pilot and feasibility study will be applied to the pilot European Examination in MM, to test the methods and processes and support judgements regarding the progression to the first European Examination in MM. The analyses will be used to propose criteria or a tool to evaluate a pilot or feasibility study in assessment.

Complex interventions guidance includes assessment of feasibility in its framework for developing and evaluating complex interventions. The introduction of a pan European Examination shares some similarities with the introduction of a complex intervention (Skivington, Matthews et al. 2021) and could align to the phases and core elements of complex intervention research; intervention, feasibility, evaluation and implementation (Skivington, Matthews et al. 2021). The Feasibility study may be designed to assess the intervention itself, looking at content, delivery, acceptability, likely cost effectiveness or capacity of providers to deliver the intervention, all of which are important in a high-stakes exam. The recent revision of the framework also focusses on the need to understand how interventions bring about change. In the context of the intervention, the pilot exam, one of the anticipated outcomes is progression towards harmonisation of training and practice of Clinical Microbiology in Europe.

4.2 Aims

1. Development of an assessment tool to evaluate the trainee experience of the European Exam in MM.
2. To assess the pilot examination against the criteria used in a pilot and feasibility study, modifying these criteria as appropriate, to create a tool to assess a pilot examination.
3. Evaluation of the exam feasibility and assessment (considering points 1 & 2 above) to see if modifications may be required for implementation of a pan European Assessment in Clinical Microbiology with respect to training and scope of practice in Europe.

4.3 Methods

The Pilot exam- a study of feasibility

4.3.1 Survey of the candidates' experience

Candidates sitting the pilot exam were asked to complete a questionnaire exploring their experience of the exam process and exam delivery platform, their view of the exam questions and their view of the overall standard of the exam. The candidate survey was completed on the same day as the exam. The survey was created on SurveyMonkey, including a formal consent form, and then had 32 further questions.

4.3.2 Evaluation of the exam feasibility - process

4.3.2.1 The four reasons for conducting a pilot study.

The reasons for conducting a pilot study were described by Thabane et al; process, resources, management and scientific (Thabane, Ma et al. 2010). These were modified to outline the reasons for a pilot examination, and the pilot exam in MM was assessed against the modified reasons.

4.3.2.2 The criteria for success of a pilot study

These were also described by Thabane et al, and were assessed in the context of the pilot examination (Thabane, Ma et al. 2010). The outcome of a pilot study can be one of the following:

- Stop - main study not feasible.
- Continue, but modify protocol - feasible with modifications.
- Continue without modifications but monitor closely - feasible with close monitoring.
- Continue without modifications - feasible as is.

4.3.2.3 The ADePT Model.

This is A process for Decision-making after Pilot and Feasibility trials (Bugge, Williams et al. 2013) was adapted to address problems identified in the pilot exam. The ADePT Model is a process for Decision-making after Pilot and Feasibility trials has three steps:

1. Deciding on the type of problem experienced and evidencing each problem identified.
    a. Issue is likely to only be a problem for the pilot.
    b. Issue is likely to be a problem for the pilot and real exam.
    c. Issue is likely to be a problem for the real exam.
2. Considering a potential range of solutions and justifying them
3. Assessment of the best options available
    a. Could the solution be effective in the real exam?

Problems identified during the pilot exam were assessed using this model and a solution agreed.

4.3.2.4 The recommendations for reporting the results of pilot studies.

Again, described by (Thabane, Ma et al. 2010), and reviewed and modified in the context of a pilot examination. The pilot exam, as a feasibility study, was assessed against the modified recommendations.

4.3.2.5 The phases and core elements of complex intervention research

The phases and core elements of complex intervention research; intervention, feasibility, evaluation and implementation (Skivington, Matthews et al. 2021), were assessed in the context of the pilot examination. The MRC framework (Chapter 1, section 1.2.2.2) was extended to focus more attention on understanding how an intervention brings about change.

4.3.3 Evaluating evidence of harmonisation of training and practice.

One of the key desired changes that is driving the intervention, the introduction of the European Exam, is to progress the harmonisation of training and practice in Europe. This change was assessed by means of two surveys and an additional analysis of the pilot exam results:

4.3.3.1 A survey looking at training is Europe

The survey, created on SurveyMonkey had fifty-four questions and was circulated to full, associate and observer members of the UEMS section MM, with reminders when necessary (Appendix 6). Participants were sent a participant information leaflet, which they acknowledged receipt of before commencing the survey. Participants completed a consent form prior to completing the survey. Ethics approval was granted by the Southeast Regional Ethics Committee, Ireland.

The questionnaire was anonymous, but each participant was asked to identify the country in which they were working. The questionnaire was divided into two sections. The first section related to the training in CM. The information collected included country of training, requirements to enter specialist training, duration of training, number of trainees and mandatory training activities. Participants were asked about the structure and content of specialist MM training in their country including how their training activities were recorded and assessed. The second section of the questionnaire looked for information on the training body, training certifier, national authority for allocation of training resources, including funding and the national specialist society.

4.3.3.2 A survey looking at scope of practice in Europe

A survey looking at scope of practice in Europe was circulated to full, associate and observer members of the UEMS section MM board, on two occasions three years apart. The survey was circulated in 2019 and again in 2022, with reminders when necessary. Participants were sent a participant information leaflet, which they acknowledged receipt of before commencing the survey. Participants completed a consent form prior to completing the survey. Ethics approval was granted by the Southeast Regional Ethics Committee, Ireland.

The survey was created on SurveyMonkey and had 69 questions concerning training and current scope of practice and used a Likert scale to determine if specific tasks were the remit of the MM; 'always', >75%, 50-75%, 25-50% and 1-25% of the time or 'never'

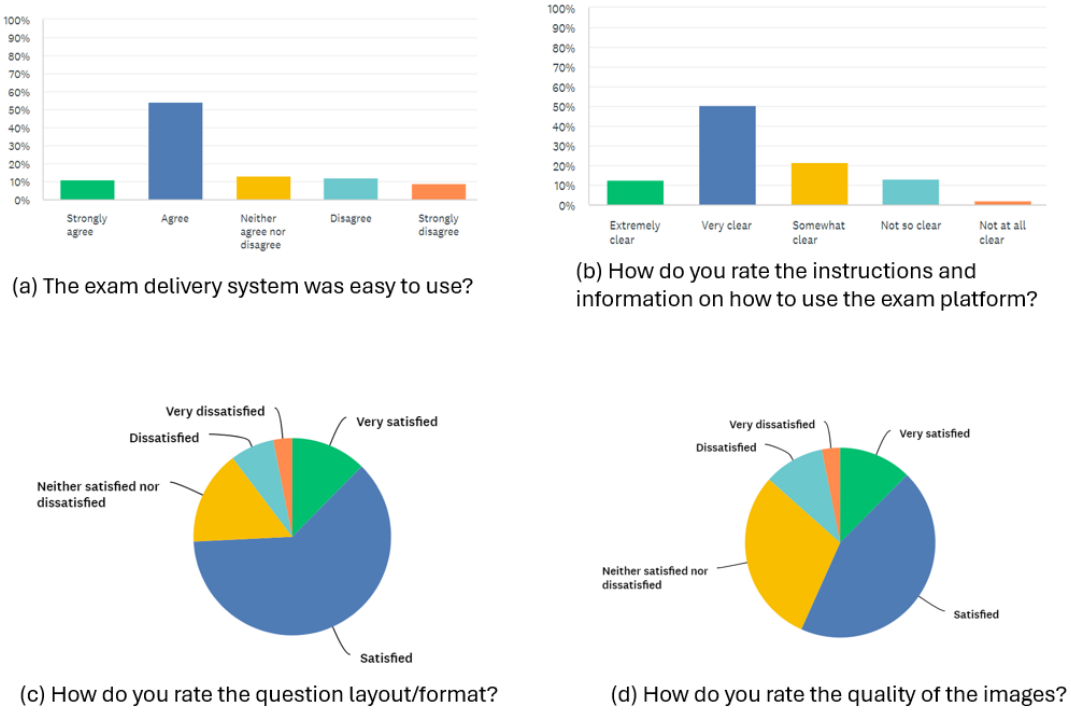
The questionnaire was anonymous, but each participant was asked to identify the country in which they were working.

4.4 Results:

4.4.1 Survey of the candidates' experience

- Ninety-six candidates from 18 UEMS affiliated countries sat the pilot exam
- 100 % of candidates completed the questionnaire
- 85 % found the registration process easy

Candidates' experiences of the examination delivery system, question layout and images are shown in Figure 4.1



(a) The exam delivery system was easy to use?

(b) How do you rate the instructions and information on how to use the exam platform?

(c) How do you rate the question layout/format?

(d) How do you rate the quality of the images?

Figure 4.1 Candidates' experiences of the examination delivery system

Candidates were asked to rate their experience of using the exam platform. Most candidates reported that the exam system easy to use (a), the instructions were clear (b) and were satisfied with the question layout (C). Over 50% reported that they were satisfied with the image quality (d).

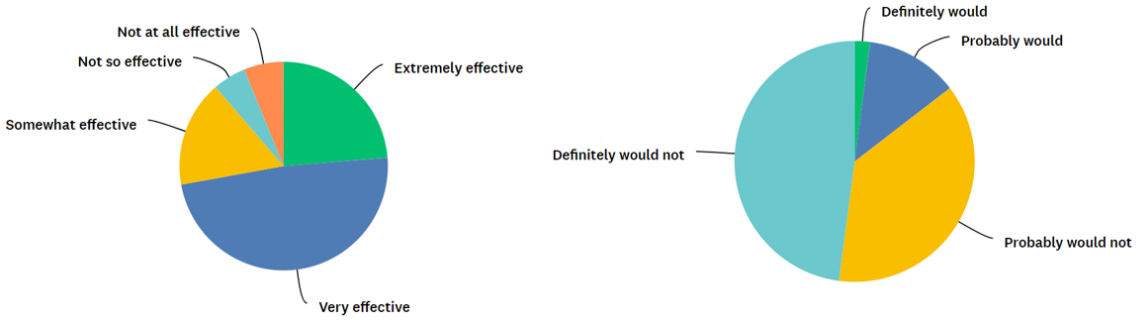
Candidates were asked whether the exam content assessed the knowledge required to become a specialist in microbiology and whether the exam covered the nine curriculum areas adequately (Table 4.1).

Question	Strongly agree/agree
The exam assessed the knowledge required to become a specialist in Medical Microbiology?	80%
Knowledge of General Microbiology was tested adequately?	76%
Knowledge of Bacteriology was tested adequately?	72%
Knowledge of Virology was tested?	81%
Knowledge of Parasitology was tested adequately?	74%
Knowledge of Mycology was tested adequately?	61%
Knowledge of Clinical Medicine was tested adequately?	73%
Knowledge of Infection Prevention and Control was tested adequately?	76%
Knowledge of Antimicrobials was tested adequately?	66%
Knowledge of Laboratory management was tested adequately?	61%

Table 4.1 Candidates were asked whether they perceived the exam assessed the knowledge required to become a specialist in Medical Microbiology and whether it assessed, in their opinion, each of the sections of the European curriculum of Medical Microbiology adequately.

80% of candidates reported that the exam assessed the knowledge required to become a specialist in MM. Candidates were less satisfied with the areas of mycology and laboratory management.

Candidates were asked about their experience of the remote invigilation and whether they thought it would be possible to cheat in the exam (Figure 4.2).



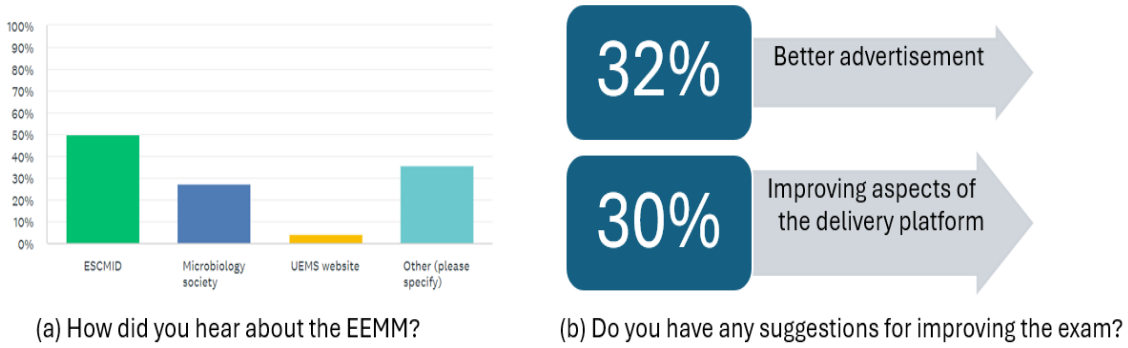
(a) What did you think of the supervision and invigilation?

(b) Do you think it would be possible for a candidate to cheat in the exam?

Figure 4.2 The pilot exam: Candidates experience of remote proctoring

Graph (a) and (b) Show that candidates thought the supervision was very effective and most reported that it would not be possible to cheat.

Candidates were asked how they heard about the exam and were asked what improvements could be made to the exam in the future. The responses are in Figure 4.3.



(a) How did you hear about the EEMM?

(b) Do you have any suggestions for improving the exam?

Figure 4.3 The pilot exam: Communication and areas for improvement.

(a) Shows how the exam was communicated to candidates. The most common source was the European Society for Clinical Microbiology and Infectious Diseases (ESCMID). (b) Shows the areas highlighted by candidates that could be improved.

4.4.2 Evaluation of the exam feasibility - process

4.4.2.1 The four reasons for conducting a pilot study

The four reasons for conducting a pilot study, described by Thabane et al, process, resources, management and scientific (Thabane, Ma et al. 2010), were adapted to outline the reasons for a pilot examination (Table 4.2).

Reasons	Description from Thabane *et al*	Adapted for the pilot exam	Examples in Thabane *et al*	Adapted to give examples in a pilot exam.
Process	*This assesses the feasibility of the processes that are key to the success of the main study*	This assesses the feasibility of the processes that are key to the success of the exam	Recruitment rates Retention rates Refusal rates Failure/success rates (Non)compliance or adherence rates eligibility criteria - Is it obvious who meets and who does not meet the eligibility requirements? - Are the eligibility criteria sufficient or too restrictive? Understanding of study questionnaires or data collection tools: - Do subjects provide no answer, multiple answers, qualified answers, or unanticipated answers to study questions?	Recruitment of candidates Exam announcement The application process Eligibility criteria - Is it obvious who meets the eligibility criteria? - Are the eligibility criteria too restrictive? - How is the candidate eligibility confirmed? Retention rates – did all the candidates that applied sit the exam? Exam fees - How will the exam fee be collected? - Will there be different exam fees for different candidates e.g. reduced fee for low-income countries? If yes, how will this be achieved?

Resources	*This deals with assessing time and resource problems that can occur during the main study*	This deals with assessing time and resource problems that can occur during the real exam	Length of time to fill out all the study forms
			
			Determining capacity:
			- Will the study participants overload your phone lines or overflow your waiting room?
			
			Determining process time
			- How much time does it take to mail out a thousand surveys?
			
			Is the equipment readily available when and where it is needed?
			
			What happens when it breaks down or gets stolen?
			
			Can the software used for capturing data read and understand the data?
			
			Determining centre willingness and capacity
			
			- Do the centres do what they committed to doing? - Do investigators have the time to Perform the tasks they committed to doing?
			- Are there any capacity issues at each participating centre?
				Length of time to sit the exam
				
				The exam delivery
				- Instructions for candidate?
				- Is the question layout clear?
				- Are the images/video clear?
				
				Determining the capacity of the exam
				- How many candidates in a sitting?
				
				Exam development resources
				- Question-writers Number? Mix? Time?
				- Question reviewers Number? Mix? Time?
				- Maintaining the database
				
				Facility and delivery
				- Face to Face exam
				  - Venue, facilities ? Toilets ? Registration ? computers, tablets, invigilators
				- Online exam
				  - Internet capacity, invigilation, security checks
				
				The exam platform security
				- Level of access for question-writers, reviewers, chair

Management	*This covers potential human and data management problems*	This covers potential human and database management problems	What challenges do study personnel have? Is there enough room on the data collection form for all of the data you receive? Are there any problems entering data into the computer? Can data coming from different sources be matched? Were any important data values forgotten about? Do data show too much or too little variability?	Question database - Overview of question numbers and question content? - tracking of questions reviewed? Question-writing - Training the question writers Question review - Internal review - External review Question selection Standard-setting - Choosing the method - training Quality measures - reliability, validity
Scientific	*This deals with the assessment of treatment safety, dose, response, effect and variance of the effect*	This deals with the assessment of the pilot exam quality and process and the evaluation of the exam	Is it safe to use the study drug/intervention? What is the safe dose level? Do patients respond to the drug? What is the estimate of the treatment effect?	Results - Pass rate of candidates sitting the pilot Evaluation of - Exam quality and process - Exam feasibility

Table 4.2 Reasons to conduct a pilot study adapted to assess reasons to conduct a pilot exam.

Table is a modified version of Table 2, in “Reasons for Conducting Pilot Studies” published in Thabane *et al* (Thabane, Ma et al. 2010). The questions have been modified to reflect the reasons for doing a pilot exam.

The modified tool to assess the reasons for conducting a pilot study was used to assess the pilot exam in MM. The findings of the assessment are summarised in Table 4.3.

Reason	Reason in a pilot exam	Examples in a pilot exam	Why?	Assessment of the pilot exam in MM against these criteria
Process	*This assesses the feasibility of the processes that are key to the success of the main study*	Recruitment	In a pilot exam, the recruitment rates are important to assess the potential interest in the exam, whether there will be candidates interested in sitting the exam. to ensure that sufficient numbers sit the exam to facilitate standard setting and psychometrics	The number and geographical spread of candidates demonstrated good interest in the exam. The recruitment of 96 candidates plus one scrutineer gave sufficient numbers to judge the psychometrics
		Exam announcement	How will the occurrence of the exam be communicated to potential candidates? Was the announcement successful in reaching candidates?	This was assessed by means of a candidate questionnaire.
			When should the announcement for the exam go out? How much time should candidates have to apply before the application closes? How soon after the application will candidates know if they are accepted?	A communication strategy and timeline for communication was developed by the exam working group.

		The application process	How easy is the application process for candidates?	This was assessed by means of a candidate questionnaire.
		Is it obvious who meets the eligibility criteria?	A process should be in place to assess applicants against the eligibility criteria	In our exam a two-step process was created to facilitate the assessment of eligibility of the candidates. The candidates filled in an application form which asked questions rating to their eligibility e.g. duration of training and certification from two supervisors that they had the required specialist training in Clinical Microbiology. The application form was assessed by a non-medical administrator and if deemed complete and eligible, was assessed by a member of the project team.
		Eligibility criteria Are the eligibility criteria too restrictive?	The eligibility criteria should be appropriate to the exam standard.	Our exam was an 'exit' exam, the eligibility criteria included a minimum duration of training. The performance of candidates indicated that this was appropriate.
		Retention rates	Did the applicants that registered to sit the pilot exam complete the exam?	The pilot exam in Microbiology had 99 applicants registered, of which 97 completed the exam (96 candidates and 1 scrutineer) Two candidates who registered did not complete the exam due to internet instability.
		Exam fees	How will the exam fee be collected? How will the exam fee be determined?	In the pilot exam, the fee was waivered. To test the fee payment mechanism, eligible candidates were issued with a unique code/exam number to verify their eligibility for the online pilot exam.

Resources	*This deals with assessing time and resource problems that can occur during the main study*	Length of time given to sit the exam	Is the time allocated for the exam appropriate for the number of questions and assessment method?	Our assessment was an MCQ format with 99 questions. The candidates had 150 to complete the assessment
		Instructions	Are the instructions for the candidate clear?	This was assessed by means of a candidate questionnaire.
		The exam delivery	Is the question layout clear? and easy for the candidate to understand?	This was assessed by means of a candidate questionnaire.
			If the exam includes images or videos, are they of sufficient quality?	This was assessed by means of a candidate questionnaire.
			How many candidates can sit the exam at one time?	In the pilot we aimed for a 50-100 to ensure robust psychometrics (Nunnally and Bernstein 1994) and to facilitate the remote invigilation.
		Determining capacity	Question-writers Number? Mix? Time?	Question-writers were recruited through the UEMS section MM. Question-writers were divided into groups based on their area of preference/expertise. Each group had a chairperson and was assigned an area of the curriculum. A target number of questions in each curriculum area was determined by the blueprint. See chapter 2.
		Exam development resources	Question reviewers Number? Mix? Time?	Reviewers were recruited through the UEMS section MM.
			Maintaining the question database	The database was maintained by the project group administrator and chairperson of the exam working group
			If the exam will be delivered face-face, is there a venue? What is the capacity?	N/A in our pilot

		Facility and delivery Face to Face exam	Does the venue have toilet facilities? Tea/coffee facilities? A waiting area?	N/A in our pilot
			Does the venue have computers or tablets?	N/A in our pilot
			How many invigilators will be needed?	N/A in our pilot
			What will happen if there is equipment failure?	N/A in our pilot
			Who will register the candidates when they arrive?	N/A in our pilot
			What internet power will the candidate need?	The exam platform company provided this information
		Facility and delivery Online exam	Is it possible for the candidate to check that their internet power is adequate to sit the exam, before the exam?	A test option was offered to candidates
			If the exam is delivered online, what type of remote invigilation will be used?	Remote invigilation was carried out by people
			What is the ratio of invigilators to candidate? Who will do the invigilating?	Each invigilator monitored 6-8 candidates. The exam project group and members of the question-writing group invigilated. The company providing the computer platform also invigilated.
			What security checks will be taken?	A series of agreed security checks were in place
			What happens if a candidates internet link fails?	Two candidates lost their internet and were unable to complete the pilot exam.

		Equipment	Access to the platform? Level of access for writers? Reviewers? Chair?	This was included in the exam platform tender document
Management	This covers potential human and database management problems	The exam platform security	Tracking of question review question numbers and content	This was included in the exam platform tender document
		The question database	Training the question-writers	Workshops were held to train the question-writers.
		Question-writing	Who will review the questions for quality and content? Internal review? External review?	Internal review occurred within the question writing group. External review was carried out by subject matter experts who were not question-writers in the group.
		Question review	How will the exam questions be selected	The exam questions were selected using the exam blueprint.
		Question selection	What standard setting method will be used?	Two standard setting methods were used and compared, see chapter 3.
		Standard-setting	Reliability Validity Content	Reliability and validity were assessed using psychometrics. Exam content was assessed by a candidate questionnaire
Scientific	*This deals with the pilot exam quality and process and evaluation of the exam*	Quality measures	Failure/Success rates of candidates sitting the pilot.	This will allow the exam team to assess whether the performance standard is appropriate for the 'level' of the candidates
		Results	Evaluation of the exam quality	This was assess using the tool described by Bartman *et al* (Baartman, Bastiaens et al. 2007) and is discussed below
		Evaluation	Evaluation of feasibility	This was done using this tool/checklist

Table 4.3: Reasons for conducting a pilot exam as a tool/checklist for the pilot examination in Medical Microbiology

The reasons for the pilot exam are assessed against the modification of Thabane's reasons for a pilot study (Thabane, Ma et al. 2010)

4.4.2.2 The criteria for success of a pilot study

The criteria for success described by Thabane, were assessed in the context of the pilot examination (Thabane, Ma et al. 2010). The outcome of a pilot study can be one of the following:

- Stop - main study not feasible.
- Continue, but modify protocol - feasible with modifications.
- Continue without modifications but monitor closely - feasible with close monitoring.
- Continue without modifications - feasible as is.

4.4.2.3 The ADePT Model.

The ADePT Model is a process for decision-making after pilot and feasibility trials (Bugge, Williams et al. 2013) and was adapted to address problems identified in the pilot exam. The ADePT Model process for decision-making after pilot and feasibility trials was applied to two problems identified in the pilot exam.

Problem	Step 1 Identify the type of problem	Step 2 Consider the solutions	Step 3 Assessment of the best solution
Remote Invigilation. The experience of delivering the exam online by remote invigilation created some concerns for the Exam Working Group around security.	Type B: Issue is likely to be a problem for the pilot and real exam	Hold the exam face-to-face. This is more expensive for both the candidates and the exam team but will address the concerns re remote invigilation	Hold the exam face to face for the first real exam. Assess the impact of the face-to-face exam on recruitment of candidates and cost of exam delivery. Explore other options for online delivery and emote invigilation. Review the literature from the pandemic to assess other exam groups experience and solutions
		Explore other options for remote invigilation, drawing on experience of other exam groups. This may lead to a delay in delivery of the real exam and risk loss of momentum for question-writers	
Administrative support for communications, the application process, uploading questions to the question bank was inconsistent and led to a lot of administrative tasks being undertaken by members of the exam working group	Type B: Issue is likely to be a problem for the pilot and real exam	Members of the working group provide administrative support. The Exam Working Group members do this work in addition to their routine clinical work, so this would be very difficult	New administrative support will be sought. Different solutions for sourcing this support will be considered. This issue will be addressed with urgency to ensure the new support can be trained.
		Recruitment of alternative administrative support. This will involve training the new administrative support to ensure they are familiar with the exam process and timelines and are able to use the exam platform	

Table 4.4: ADePT Model process for Decision-making

The adaptation of the ADePT Model process for Decision-making was applied to two problems identified in the pilot exam. The first problem was the experience of remote invigilation. The second problem was administrative support.

4.3.2.4 The recommendations for reporting the results of pilot studies.

A checklist for reporting the results of a pilot study is outlined in Table 3 of Thabane's paper, "A tutorial on pilot studies: the what, why and how" (Thabane, Ma et al. 2010), The checklist criteria were modified to create a tool for reporting a pilot exam (Table 4.5).

PAPER SECTION	Item	Descriptor in pilot study	Descriptor in pilot exam
TITLE and ABSTRACT	1	Does the title or abstract indicate that the study is a "pilot"?	Does the title or abstract indicate that the study is a "pilot"?
Background	2	Scientific background for the main study and explanation of rationale for assessing feasibility through piloting	Scientific background for the development of the exam and explanation of rationale for assessing feasibility through piloting.
Participants and setting	3	Eligibility criteria for participants in the pilot study (these should be the same as in the main study, if different, state the differences) The settings and locations of data collection	Eligibility criteria for candidates in the pilot exam (these should be the same as in the main exam – if different, state the differences) e.g. communication, the application process
Interventions	4	Provide precise details of the interventions intended for each group and how and when they were administered (if applicable) – state clearly if any aspects of the intervention are assessed for feasibility	Provide precise details of the exam Development process, including question-writing and quality assurance, the Exam delivery process, the platform, the question database – state clearly if any aspects of the intervention are assessed for feasibility
Objectives	5	Specific scientific objectives and hypotheses	Specific objectives and reason for the development of the exam

Outcomes	6	Specific feasibility objectives Clearly defined primary and secondary outcome measures for the main study Clearly define the feasibility outcomes and how they were operationalised – including key elements such as recruitment rates, consent rates, completion rates, variance estimates, etc	Specific feasibility objectives e.g. quality question writing and review, question-bank management. Clearly defined primary and secondary outcome measures for the exam e.g. develop a quality exam. Clearly define the feasibility outcomes and how they were operationalised – these should include key elements such as standard setting
Sample size	7	Describe how sample size was determined.	Describe how sample size was determined to ensure validity and reliability can be assessed. What is the minimum number of candidates to assess feasibility? What is the maximum e.g. to facilitate invigilation or if face-to-face, the venue may limit numbers.
Feasibility Criteria	8	Clearly describe the criteria for assessing success of feasibility – these should be based on the feasibility objectives	Clearly describe the criteria for assessing success of feasibility e.g. test reliability, validity, candidate feedback
Statistical Methods	9	Describe the statistical methods for the analysis of primary and secondary feasibility outcomes	Describe the methods that will be used to assess quality, validity and reliability.

Ethical Aspects	10	State whether the study received research ethics approval. State how informed consent was handled – given the feasibility nature of the study	State whether the pilot exam or any of its processes received research ethics approval. State how informed consent was handled. – given the feasibility nature of the study
Participant flow	11	Flow of participants through each stage (a flow-chart is strongly recommended). Describe protocol deviations from pilot study as planned, together with reasons. State the number of exclusions at each stage and reasons for exclusions	Flow of participants through each stage (a flow-chart is strongly recommended). Describe the process from announcement to application, to verifying eligibility, exam delivery and communication of results
Recruitment	12	Report the dates defining the periods of recruitment and follow-up	Report the dates of the communication, applications opening and closing, date of exam and date of results
Baseline data	13	Report the baseline demographic and clinical characteristics of the participants	Describe the eligibility criteria
Outcomes and estimation	14	For each primary and secondary feasibility outcome, report the point estimate of effect and its precision (e.g., 95% confidence interval [CI]) – if applicable	Review the exam results, psychometric analysis, candidate feedback

Interpretation	15	Interpretation of the results should focus on feasibility, taking into account the stated criteria for success of feasibility; study hypotheses, sources of potential bias or imprecision	Interpretation of the results should focus on feasibility, taking into account the stated criteria for success of feasibility. e.g. evidence of quality, such as content validity, construct validity, internal validity, external validity
Generalisability	16	Generalisability (external validity) of the feasibility. State clearly what modifications in the design of the main study (if any) would be necessary to make it feasible	Generalisability: Can the pilot exam processes be applied to the real exam? Will parts of the process need modification
Overall evidence of feasibility	17	General interpretation of the results in the context of current evidence of feasibility Focus should be on feasibility	General interpretation of the results in the context of current evidence of feasibility

Table 4.5 Pilot Study - Checklist: Items to include when reporting a pilot study, adapted to facilitate reporting of a pilot exam.

The checklist by Thabane *et al* (Thabane, Ma et al. 2010), for reporting a pilot study was adapted to create a checklist for reporting a pilot exam. Thabane's checklist is in column 1-3. The adaptation for a pilot exam is in the fourth column.

4.4.2.5 The phases and core elements of complex intervention research

Six core elements of complex intervention research

The phases and core elements of complex intervention research; intervention, feasibility, evaluation and implementation (Skivington, Matthews et al. 2021), were assessed in the context of the pilot examination. Six core elements should be considered and are summarised in Table 4.6.

Core elements	**The pilot Examination in Medical Microbiology**
How does the intervention interact with its context?	The UEMS vision is harmonisation of training and practice in Europe. Harmonisation of training and practice will facilitate movement of doctors within the EU.
What is the underpinning programme theory?	The European exam is blueprinted against the European Curriculum. The exam tests in the knowledge domain in all areas of the curriculum. The scope of practice in clinical microbiology may vary from country to country as demonstrated in the scope of practice surveys. Assessment drives learning.
How can diverse stakeholder perspectives be included in the research?	Diverse stakeholders were included in a number of ways: • The Exam Working Group was made up of members of the UEMS section MM. • Members of the UEMS section of MM were invited to participate in the exam development, as question writers and reviewers. • The exam received support from ESCMID • Exam candidates were invited to give feedback on their experience of and content of the pilot exam.

What are the key uncertainties?	What number of candidates are likely to sit the exam each year? Will there be sufficient administrative support to sustain the exam delivery? As a small section, will there be enough interest from the section members and their colleagues to sustain the question writing and review.
How can the intervention be refined?	Challenges identified during the pilot exam need to be assessed and solutions sought in order to successfully deliver the real exam. This will include engagement with the candidates, through the candidate feedback survey.
What are the comparative resource and outcome consequences of the intervention?	Challenges and decisions around resources, human and financial are considered in the ADePT model for decision making, section 4.3.2.3 The regular delivery of a sustainable, high quality, EEMM that is affordable for candidates is essential to develop the face validity of the exam, to achieve the outcome of harmonizing scope of practice and training in Europe.

Table 4.6: The six phases of complex intervention research

The core elements of complex intervention research were considered in the context of the pilot examination.

Phases of Complex Intervention Research.

The pilot exam is assessed against the phases of complex intervention research in Table 4.7.

The phases of complex intervention research	**The pilot Examination in Medical Microbiology**
Identification of the intervention	The European Exam in MM.
Feasibility	The pilot exam
Evaluation	This chapter in the thesis
Implementation	The first European Examination in MM was held 14 months after the pilot

Table 4.7 The core elements of complex intervention research

The core elements of complex intervention research were considered in the context of the pilot examination.

Proposed tool to assess a pilot exam.

A tool was developed incorporating elements of Thabane's tools (Thabane, Ma et al. 2010) and elements of the complex intervention tool (Skivington, Matthews et al. 2021) to assist researchers in developing and assessment of a pilot exam (Table 4.8).

Objectives	Specific objectives and reason for developing the pilot exam.	
Outcomes	Clearly defined primary and secondary outcome measures for the pilot exam	
Reasons to do a pilot exam	**The pilot exam**	**Examples in a pilot exam**
Process	This assesses the feasibility of the processes that are key to the success of the exam	Recruitment of candidates Exam announcement The application process Eligibility criteria - Is it obvious who meets the eligibility criteria? - Are the eligibility criteria too restrictive? - How is the candidate eligibility confirmed? Retention rates – did all the candidates that applied sit the exam? Exam fees - How will the exam fee be collected? - Will there be different exam fees for different candidates e.g. reduced fee for low-income countries? If yes, how will this be achieved?

Resources	This deals with assessing time and resource problems that can occur during the real exam. -exam delivery	Length of time to sit the exam The exam delivery - Instructions for candidates? - Is the question layout clear? - Are the images/video clear? Determining the capacity of the exam - How many candidates in a sitting? Exam development resources - Question-writers Number? Mix? Time? - Question reviewers Number? Mix? Time? - Maintaining the database Facility and delivery - Face to Face exam - Venue, Facilities? Toilets? Registration? Computers, tablets, invigilators - Online exam - Internet capacity, invigilation, security checks The exam platform security - Level of access for question-writers, reviewers, chair

Management	This covers potential human and database management problems	Question database - Overview of question numbers and question content? - tracking of questions reviewed? Question-writing - Training the question writers Question review - Internal review - External review Question selection Standard-setting - Choosing the method - training Quality measures - reliability - validity
Scientific	This deals with the assessment of the pilot exam quality and process and the evaluation of the exam	Results - Pass rate of candidates sitting the pilot Evaluation of - Exam quality and process - Exam feasibility

Feasibility criteria	Clearly describe the criteria for assessing success of feasibility e.g. test reliability, validity, candidate feedback, acceptability, likelihood of cost effectiveness and capacity of providers to deliver the intervention, the exam.
The outcome of the pilot exam	• Stop - exam not feasible. • Continue, but modify elements of the process - feasible with modifications. • Continue without modifications but monitor closely - feasible with close monitoring. • Continue without modifications - feasible as is.

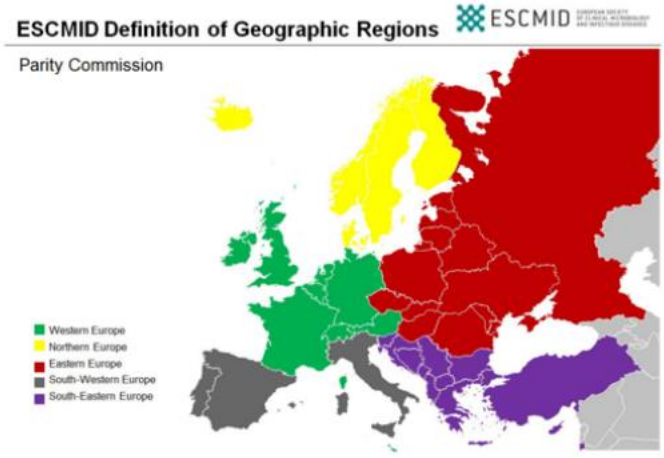
Table 4.8 Proposed tool to assess a pilot exam; to assess the feasibility of developing a new high-stakes examination.

A tool was developed incorporating elements of Thabane's tools for pilot studies (Thabane, Ma et al. 2010) and elements of the complex intervention tool (Skivington, Matthews et al. 2021) to assist researchers in development and assessment of a pilot exam (Table 4.8).

4.4.3 Evaluation: evidence for harmonisation?

4.4.3.1 A survey looking at training is Europe

A survey looking at training in Clinical Microbiology in Europe was undertaken (Doyle, Boyle et al. 2021). Responses were received from forty-four people, representing 21 countries and all Geographic areas, as defined by ESCMID (Figure 4.4).



(a) Map of Geographic regions

Geographic Region	Trainee survey	Section MM
Western Europe	9	9
Northern Europe	3	4
Eastern Europe	3	6
South-Western Europe	3	2
South-Eastern Europe	4	6

(b) Number of countries in each geographic region in the survey and in the section of Medical Microbiology

Figure 4.4 The Countries which responded to the training survey, divided into ESCMID geographic regions

The map, (a), shows the geographic areas as defined by ESCMID's parity Commission. The table, (b) shows the number of countries in the UEMS section MM in each geographic area and the number of countries that participated in the survey, broken down into each geographic area. The Trainee survey had responses from all of the five ESCMID defined Geographic areas. The highest response was from countries in Western Europe. The lowest response was from countries in Eastern Europe and South-western Europe. The column labelled section MM indicates the number of countries in each area in the UEMS section of MM.

Trainees were asked the judge the percentage of time spent in different training activities. The training activities were reported in Figure 4.5, which shows which section of the curriculum the training activities align to.

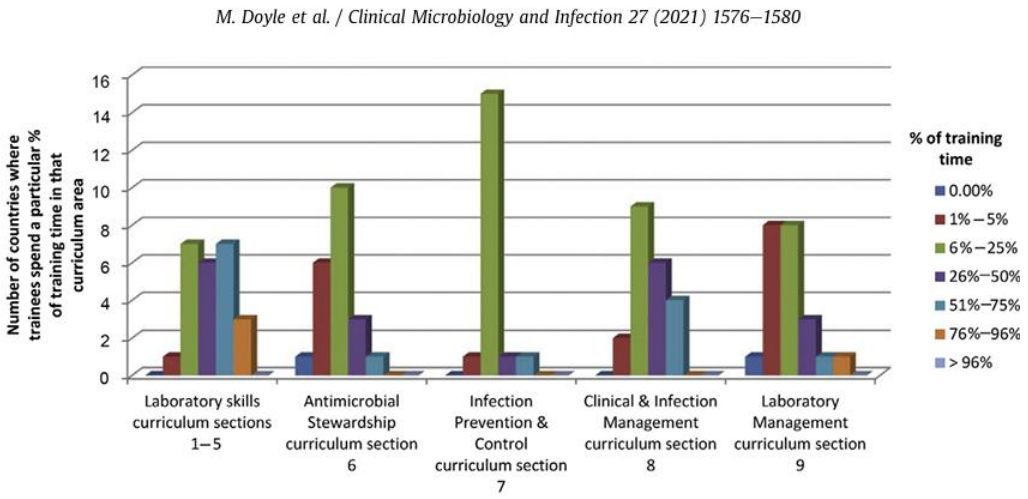


Figure 4.5 The proportion of training time dedicated to each area of the curriculum.

Percentage of trainee's time dedicated to each of the areas; Laboratory skills, Laboratory management, Clinical & infection management, Infection prevention and control, antimicrobial stewardship. The y-axis shows the number of countries. Some countries displayed an even distribution across each of the areas of the curriculum, as noted; however, in one country trainees spent no time training in Laboratory management, and in one country, no time was spent in Antimicrobial Stewardship activities. Most countries showed a good spread across all areas of required competency. In all countries, trainees continued to gain experience in Laboratory skills and 95% had experience in Laboratory management.

The questionnaire also explored the structure of the training programmes including how they align with the ETR. This is summarised in Figure 4.6., which is reproduced from Doyle *et al* (Doyle, Boyle et al. 2021).

The current structure of training programmes in clinical microbiology in Europe

Country	Do you have to be a medical doctor to train as a CM specialist in your country?	Do you have to do a formal CM training programme in order to register as a specialist?	In your country, is there a structured training programme for trainees who wish to become medical microbiology specialists?	Does the trainee have to spend time working in clinical medicine before specialist training[a]	Minimum duration in clinical medicine before specialist training	What is the duration of CM training?[a]	Are trainees required to pass a specific national or international exit exam in order to register as a specialist?	How many trainees are currently on the CM training programme	What is the average annual intake of new trainees?	Is the medical microbiology training in your country national or regional or all at one site?	Flexible training option	Logbook	Out of clinical programme experience e.g. research	Are trainees reassigned new trainers for each year of training or one for the whole programme	Is there a training curriculum with specific learning outcomes?
Austria	Yes	Yes	Yes	Yes	1 year	5 or 6 years	Yes	<25	<10	One site	Yes	Paper	Yes	One	Yes
Belgium	Yes	Yes	Yes	No		4 years	No	<25		National rotation	No	Electronic	No	Yearly change	Yes
Croatia	Yes	Yes	Yes	Yes	1 year	5 years	Yes	25–50	10–20	Regional hub	No	Paper	No	One	Yes
Czech Republic	Yes	Yes	Yes	Yes		5 years	Yes	<25	<10	Regional hub	No	Paper	No	One	Yes
Denmark	Yes	Yes	Yes	Yes	1 year	5 years	No	<25	<10	Regional hub	No	Electronic	No	Yearly change	Yes
Finland	Yes	Yes	Yes			5 years	Yes	<25	<10	Varies	Yes	Paper	Yes	Varies	Yes
Germany	Yes	Yes	Yes	Yes	1 year	5 years	Yes	25–50	10–20	One site	Yes	Paper	Yes	One	yes
Greece	Yes	Yes	Yes	No		2 years	Yes	<25	<10	One site	No	Paper	No	One	Yes
Ireland	Yes	Yes	Yes	Yes	3 years	5 years	Yes	<25	<10	National rotation	Yes	Electronic	Yes	Yearly change	Yes
Italy	Yes	Yes	Yes	No		4 years	Yes	25–50	<10	Varies	No	Paper	Yes	Varies	No
Netherlands	Yes	Yes	Yes	No		5 years	No	76–100	10–20	Regional hub	Yes	Paper	Yes	Varies	Yes
Norway	Yes	Yes	Yes	Yes	1 year	5 years	No	25–50	<10	One site	Yes	Both	Yes	One	Yes
Poland	No	Yes	Yes	No		4 years	Yes	<25	<10	Regional hub	Yes	Electronic	Yes	One	Yes
Portugal	Yes	Yes	Yes	Yes	9 months	5 years	Yes	25–50	<10	One site	Yes	Paper		One	Yes
Slovakia	No	Yes	Yes	No		4 years	Yes	<25	10–20	One site	Yes	Electronic	No	One	Yes
Slovenia	Yes	Yes	Yes	Yes	1 year	5 years	Yes	<25	<10	Varies	No	Electronic	Yes	Varies	Yes
Spain	No	Yes	Yes	No		4 years	No	>100	>50	One site	No	Paper	Yes	One	Yes
Sweden	Yes	Yes	Yes	Yes	18 months.	5 years	No	51–75	10–20	Regional hub	Yes	Paper	Yes	One	Yes
Switzerland	No	Yes	Yes	No		4 years	Yes	25–50	10–20	Varies	Yes	Paper	Yes	One	Yes
Turkey	No	Yes	Yes	No		4 years	No	>100	>50	One site	No	Electronic	Yes	One	Yes
UK	No	Yes	Yes	Yes	2 years	4 years	Yes	>100	>50 infection trainees	Regional hub	Yes	Yes	Yes	Varies	Yes

[a] Questions that relate specifically to the European training requirements.

Figure 4.6 The current structure of training programmes in Clinical Microbiology in Europe

This figure was reproduced from Doyle *et al.* (Doyle, Boyle et al. 2021). The questionnaire explored the structure of the training programmes including how they align with the European Training Requirements

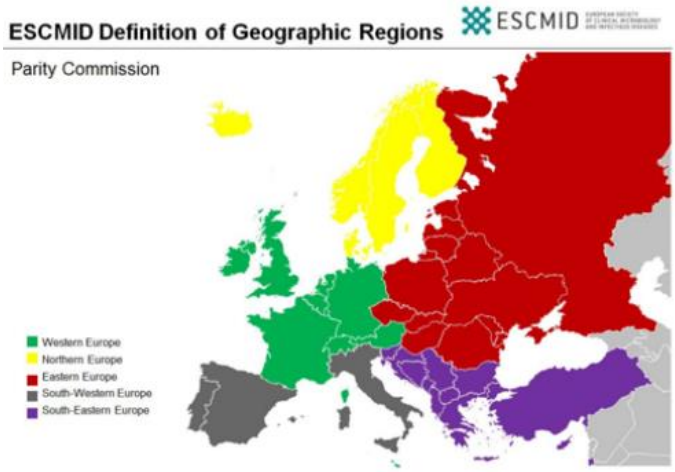
All trainees recorded their training activities on paper or electronically. The training activities recorded in the trainee's logbook/training record are listed in Figure 4.7. This figure was reproduced from Doyle *et al.* (Doyle, Boyle et al. 2021).

Area	Activity	% of countries
Laboratory experience and demonstration of practical skills	Culture (bacteria, fungi, mycobacteria)	86%
	Identification (e.g MALDI-TOF, biochemical)	90%
	Microscopy	90%
	Nucleic acid detection	86%
	Sample preparation/set up	81%
	Serological testing	81%
	Antimicrobial susceptibility testing	86%
Record of clinical liaison and infection management training activities in logbook	General practitioners	43%
	Haematology/oncology	33%
	Surgery	33%
	Medicine	38%
	Paediatrics	38%
	Obstetrics	29%
	Public health	38%
	Food and water microbiology	19%
	Unusual or complicated clinical cases	24%
	Intensive care unit ward rounds	29%
	Antimicrobial stewardship ward rounds	33%
Academic	Teach undergraduate students	29%
	Teach post-graduate students	43%
	Audit required	57%
	Publication required	48%
	Presentation at national/ international meeting required	33%
Management experience	Workload management	29%
	Laboratory accreditation	48%
	Development of practice guidelines	52%
	Laboratory management meetings	43%
Infection prevention and control	Surveillance	52%
	Outbreak management	52%
	Disinfection/sterilization	50%
	Infection prevention and control committee meetings	48%
Other	Mandatory study days	52%
	On-call or 'out-of-hours' commitment	57%
	Mandatory training courses	67%
Assessment	MiniCex—(a mini clinical experience is based on a 15-minute observation of a single interaction)	48%
	Directly observed procedure	67%
	Case-based discussion	57%
	Pass a national/international exit examination	62%

Figure 4.7 The percentage of countries that engage in specific training activities.

4.4.3.2 A survey looking at scope of practice in Europe on two occasions, with a three-year interval

In 2022, 22 respondents from 16 European countries participated, compared to 39 respondents from 16 countries in 2019.



(a) Map of Geographic regions

Geographic Region	2019	2022	Section MM
Western Europe	6	5	9
Northern Europe	4	3	4
Eastern Europe	3	2	6
South-Western Europe	1	2	2
South-Eastern Europe	3	3	6

(b) Number of countries in each geographic region in the survey in 2019 and 2022 and in the section of Medical Microbiology

Figure 4.8 The Countries which responded to the scope of practice surveys, divided into ESCMID geographic regions

The map, (a), shows the geographic areas as defined by ESCMIDs parity Commission. The table, (b) shows the number of countries in the UEMS section MM in each geographic area and the number of countries that participated in the two surveys, broken down into each geographic area. The scope of Practice surveys had responses from all of the five ESCMID defined Geographic areas. The highest response was from countries in Western Europe. The lowest response was from countries in South-western Europe. The column labelled section MM indicates the number of countries in each area in the UEMS section of MM.

Training

In 2022, expansion and harmonisation of training was shown to have occurred within 3 years with 100% reporting a National Training programme, with 95% a designated lead and 72% reporting a requirement to sit an exam to be considered a specialist, compared to 2019, when 80% reported a National training programme, 81% a designated lead and 61% reported exam requirement.

Role in the Laboratory

In laboratory practice, in 2022, respondents were doing less laboratory bench-work than reported in 2019 but expanding their role in laboratory leadership and management (see Table 4.9).

Laboratory scope of practice (a)	**2022**	**2019**
Analytical phase	**never**	**never**
Specimen set up	26.26%	16.22%
Performs the Molecular assays	43.75%	21.99%
Performs the antimicrobial susceptibility tests	38.89%	11.76%
Performs the tests for identification of organisms	38.89%	17.65%
Performs the serology assay/test	53.33%	26.47%
Performs the examination for protozoa/helminths	23.53%	9.09%
Performs the tests for the examination of skin/hair/nails	43.75%	17.65%

Laboratory scope of practice (b)	**2022**	**2019**
Post-analytical	**>75%**	**>75%**
Interpret the test results to give clinical advice	95	72.23
Interpret the test result to give IPC advice	80	63.89
Lab Management and leadership		
Developing SOPs for specimen set up, culture and microscopy	83.34	74.29
Lead role in overseeing quality in the laboratory	80	63.89
Lead role in development and review of the laboratory SOPs	84.21	69.45
Lead role in ensuring safety in the laboratory	84.21	48.57

Table 4.9 (a) and (b): Changes in the scope of practice in Clinical Microbiology in the Laboratory between 2022 and 2019

The respondents were asked if specific tasks were their role. The proportion of respondents who reported 'never' doing routine benchwork increased in 2022. The proportion of respondents undertaking post analytical tasks and leadership roles in the laboratory increased in 2022.

Role in Infection Prevention and control

In Infection Prevention and control (IPC), respondents reported more involvement in outbreak management and device-related infections in 2022, but less involvement in guideline review and development (Table 4.10).

Infection Prevention and Control practice	**2022**	**2019**
	>75%	**>75%**
Lead role in Hospital outbreaks of infection	50	44.44
Leads the Infection Prevention and Control team in the Hospital	50	41.18
Gives advice on patient isolation requirements	29.41	27.78
Gives advice on prevention of device-related infection	35.3	30.55

Table 4.10 Changes in the scope of practice in Clinical Microbiology in Infection Prevention and Control.

Respondents reported more involvement in leadership, outbreak management, and prevention of device-related infections in 2022.

Role in Clinical liaison, Antimicrobial Stewardship and teaching

Respondents reported more participation in ward rounds and multi-disciplinary (MDT) infection meetings in 2022 than in 2019 (Table 4.11). In 2022, 75% were responsible for delivery of Microbiology teaching in medical school, compared to 70.59% in 2019.

Antimicrobial Stewardship practice	2022	2019
	>75%	**>75%**
Participates in MDT/ward round with specialist services e.g. haematology	42.1	28.58
Participates in a ward round in the Critical Care Unit	35	26.47
Participates in a ward round of non-critical care patients with infection	21.05	15.15

Table 4.11: Changes in the scope of practice in Clinical Microbiology in Antimicrobial Stewardship Practice

Respondents reported more participation in ward rounds and multi-disciplinary (MDT) infection meetings in 2022 than in 2019.

4.3.3.3 Additional analysis of Exam results

The results of the pilot exam in Clinical Microbiology were reviewed looking at the examinee duration of training and the examinees country of training. Results from candidates from 5 different countries A, B, C, D and E are shown in Figure 4.9, each of which had more than 5 candidates and a group labelled 'other', with candidates from the other countries.

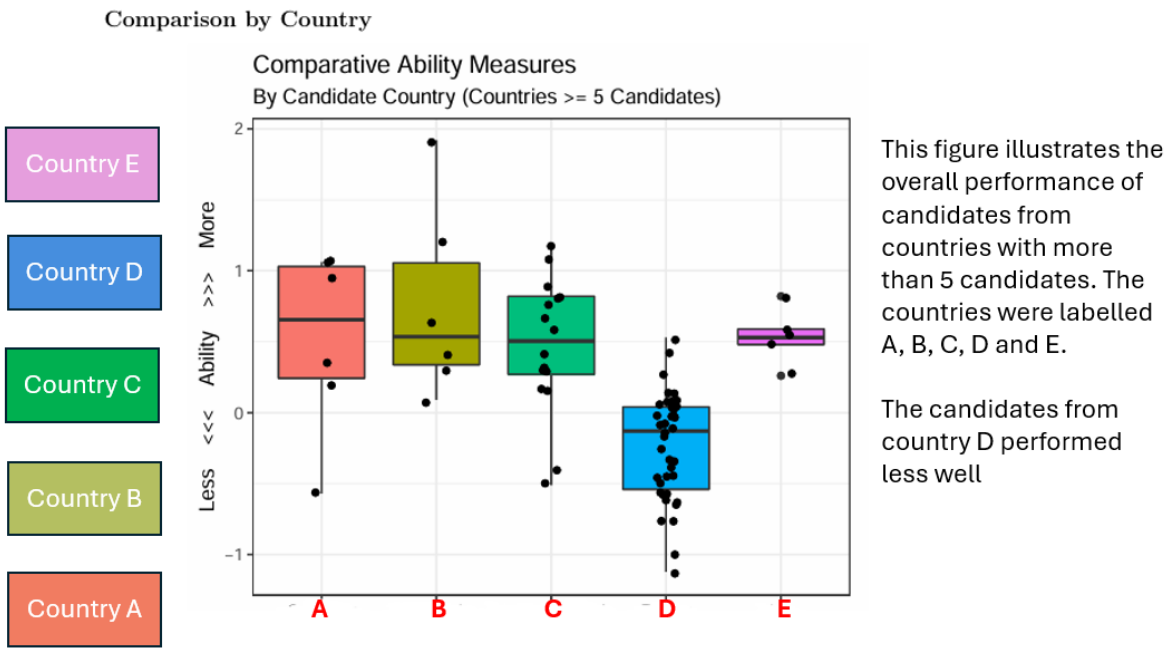


Figure 4.9 Candidate performance & Country of Training

Results from candidates from 5 different countries A, B, C, D and E are shown.

The results of trainees were compared to the results of consultants (all countries) sitting the exam (all countries). Overall, the trainees performed better than the Consultants (Figure 4.10)

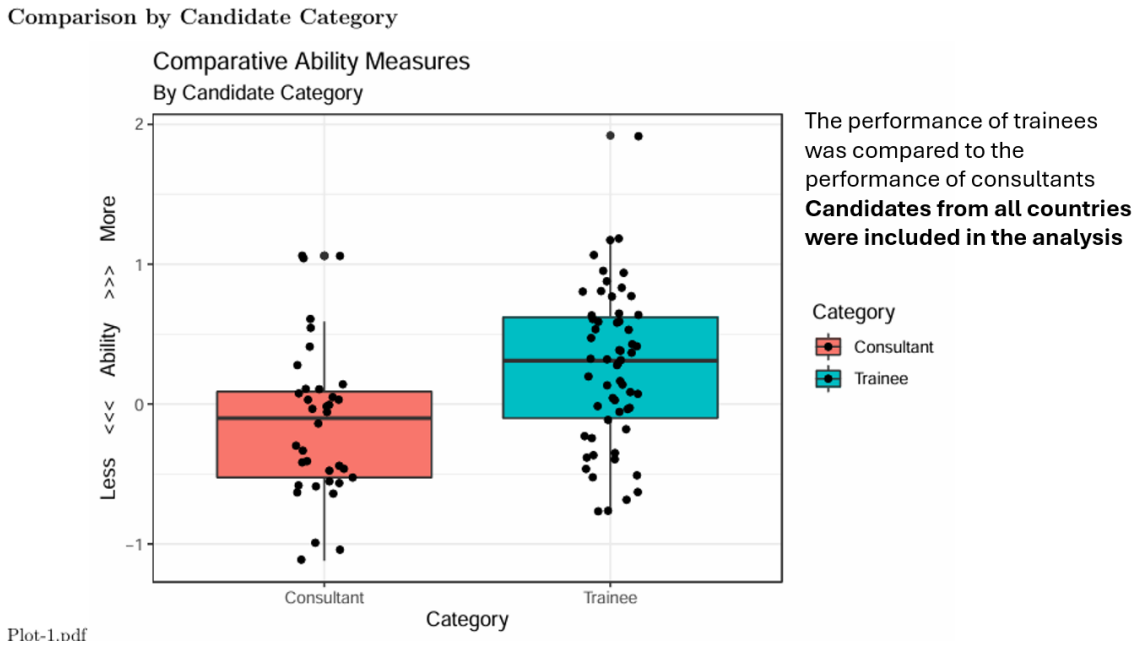


Figure 4.10 Comparison of exam candidates who were Consultants and exam candidates who were trainees

The results of trainees were compared to the results of consultants (all countries) sitting the exam (all countries). Overall, the trainees performed better than the Consultants.

The results from candidates in country D were excluded and the results of trainees were compared to the results of consultants. The difference between consultant and trainee performance were more marked (Figure 4.11).

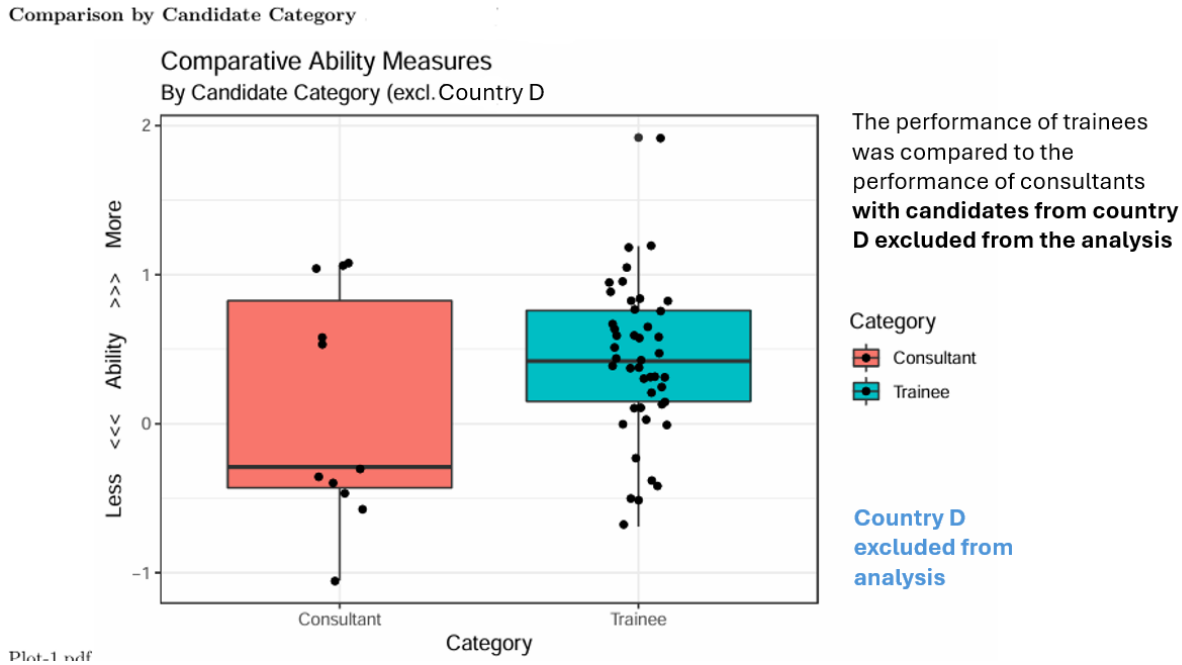


Figure 4.11 Comparison of exam candidates who were Consultants and exam candidates who were trainees, excluding country D.

The difference between consultant and trainee performance were more marked when country D candidate results were excluded.

Figure 4.12 illustrates the overall performance of candidates from countries A, B, C, D, E and other countries in the graph at the bottom right. The performance of candidates from different countries in each of the nine curriculum areas are also presented e.g. Country A performed well in most areas of the curriculum except parasitology and Country D performed less well overall but performed well in parasitology.

Figure 4.12 Performance of candidates across the 9 curriculum areas in countries A to E.

The overall performance of candidates from countries A, B, C, D, E and other countries. The performance of candidates from different countries in each of the nine curriculum areas are also presented

The candidates' performance in the exam was reviewed with respect to length of training. Overall, the relationship between duration of training and candidates' results was very weak (Figure 4.13)

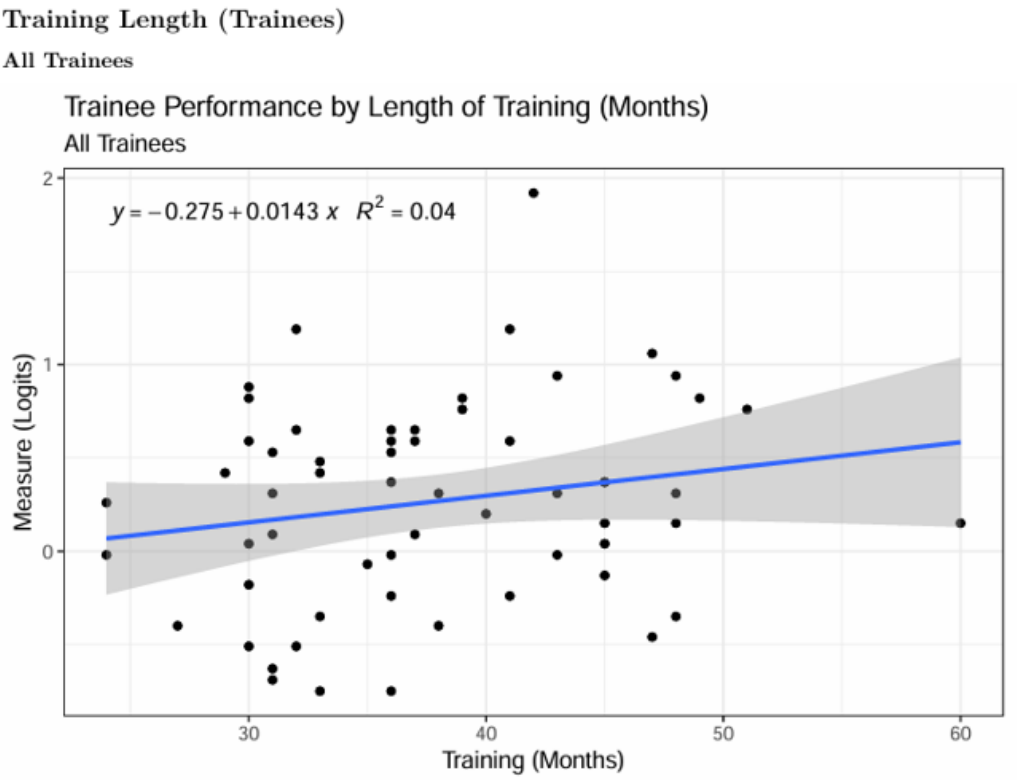


Figure 4.13 Relationship between candidate performance and duration of training

Overall, the relationship between duration of training and candidates' ability (logits) was very weak. R2 = 0.04, a 4% correlation.

The performance mapped against duration of training was reviewed for individual countries (Figure 4.14).

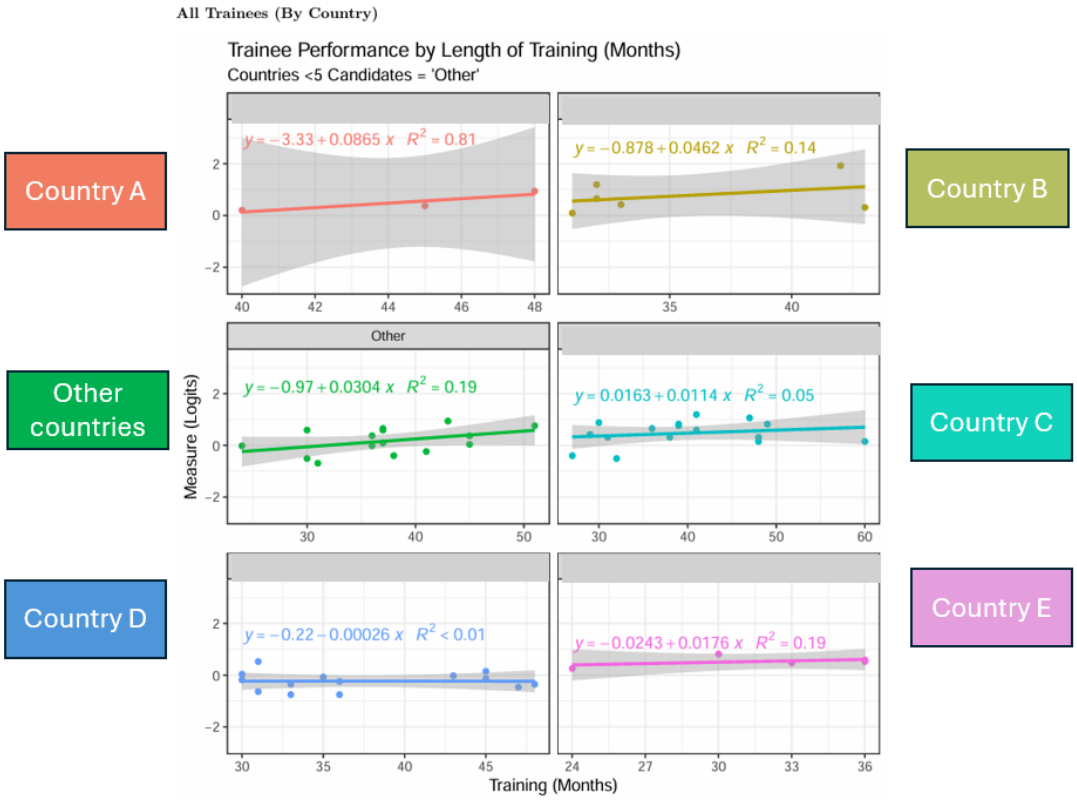


Figure 4.14 Relationship between candidate performance and duration of training in countries A, B, C, D, E and other.

This figure illustrates the overall performance of candidates from countries A, B, C, D, E and other countries. The relationship between duration of training and candidates' results was generally very weak, particularly for country D. It was stronger for country A, but there very few candidates from country A sat the exam.

Exploration of an outlier

Additional analyses explored the performance of candidates from different countries in each of the nine curriculum areas e.g. Country A performed well in most areas of the curriculum except parasitology and Country D performed less well overall but performed well in parasitology.

Given Country D's position as an outlier, the results from the surveys of scope of practice for Country D were reviewed in detail. Country D has a national training programme in CM which is 4 years duration, and the trainees sit a national exam. In 2019, the trainees did not work as a general clinical doctor prior to starting training, whereas in 2022, they worked as a general clinical doctor for 1 or 2 years prior to starting CM training. Results from the Laboratory scope of practice survey in 2019 indicated that performing many of the laboratory bench tasks (e.g. setting up specimen, performing microscopy, performing the antimicrobial susceptibility, performing serology assays) were *always* the role of the CM in 2019. In 2022, this was the CM role in 51-75% of the time but had also become the role of other laboratory staff or nursing staff. Furthermore, in Country D, the CM is always responsible for the examination of blood/stool/other tissue for the presence of protozoa/helminths and this remained their role in 2022. This may explain why examinees from Country D performed very well in parasitology. In Country D, the CM is responsible for the laboratory leadership roles e.g. making decisions around the laboratory standard operating procedures (SOPs), leading laboratory accreditation, management, quality control, safety and making decisions around introducing new technology. This was always the role of the CM in Country D in 2019 and remained their role in 2022.

The results from Country D from the 2021 survey on training in Clinical Microbiology were reviewed. In Country D, it is not necessary to be a medical doctor to train in Clinical Microbiology, however, the respondent noted that in the 8 years prior to the survey, no positions in CM were reserved for non-medical professions. MM is a primary speciality in Country D, with over one hundred trainees who complete their 4 years training in one training site. Trainees have a curriculum and a training plan and record training activities in a logbook. They record activities such as audit, publications, attendance at IPC meetings and laboratory management meetings. Participation in ICU ward rounds and antimicrobial stewardship ward rounds occurs during a three-month training period in the Infectious Diseases department. This may explain why Country D performed less well in the Clinical Medicine and Antimicrobials section of the curriculum compared to other examinees.

When compared to other countries, trainees in Country D spend more time in laboratory-focussed training, more time doing practical bench-work and get more laboratory management experience. While less time is spent in IPC activities, the trainees do attend IPC meetings and receive training in outbreak management, disinfection, sterilisation.

4.5 Discussion:

A Pilot European Exam in Medical Microbiology, a feasibility study.

The terminology regarding pilot or feasibility is the subject of debate. Some authors believe that 'feasibility' should be used when the study is more flexible in its methodology, whereas 'pilot' studies have more rigorous methodology (Arain, Campbell et al. 2010). Other authors use the terms interchangeably, that both are designed to assess aspects of the planned methods (Polit and Beck 2017) and others believe that while all pilot studies are feasibility studies, the reverse is not true (Donald 2018). Junyong In, in his 2017 paper, *introduction of a pilot study*, states in the objectives of a pilot study, that it is crucial that a pilot study validates the feasibility of the study and that it includes all the procedures for the study it was designed to pilot (In 2017). Junyong In's view on what is crucial in a pilot study are equally relevant in a pilot exam.

Researchers often use pilot studies to answer specific methods questions and to evaluate their planned procedures (Polit and Beck 2017). The pilot study can be used to test the planned methods, before applying the methods to a larger scale study (Arain, Campbell et al. 2010). While many researchers use a pilot study to answer a research question, Polit and Beck are clear that this is not the purpose of a pilot study, but instead the purpose is to provide the researchers with good knowledge of the methods (Polit and Beck 2017). Undertaking a pilot exam prior to the delivery of the first EEMM, afforded the pilot exam project group the opportunity to test all the methods and processes including the application process, blue-printing, question writing and review, standard-setting and reporting the results. The participants should be made aware that they are in a pilot study (Arain, Campbell et al. 2010). In the pilot exam, candidates were made aware that it was

a pilot exam and that there would be no award, or certificate would be made to candidates that achieved the standard.

A pilot study may also be used to test instruments for data collection and methods for handling data (Lowe 2019); this is also relevant for the pilot exam, where we were able to assess our data management instruments such as the question bank, uploading questions, using the question bank platform to facilitate reviewing and tracking the question review. This was particularly important as question writers and reviewers were working in different parts of Europe. The exam delivery platform used was the European Diploma in Radiology (EDiR) platform in collaboration with the European Board of Radiology (EBR). The feedback from the the candidate experience survey was overall very positive, with respect to the registration, instructions, and most aspects of the exam delivery. This was important information for the Exam Working Group because it was decided that the same platform will be used to deliver the first European Exam in MM. The feedback with respect to language was very reassuring. The exam is in English, but for many candidates, English is not their first language. This potentially creates challenges when writing and presenting questions that are appropriate. There was wide representation from countries throughout Europe.

The candidate experience survey reported high confidence in the exam security and effectiveness of the remote invigilation, however, the Exam Working Group had some concerns regarding the remote invigilation, which was a new experience for the group. This concern was explored using the ADePT model (Bugge, Williams et al. 2013) for decision-making after a pilot or feasibility study (Table 4.4). Following this assessment, it was agreed that the first real exam would be delivered face to face with invigilators in the room. The option for remote invigilation in future diets of the exam would continue to be

explored, through personal communications with other exam groups and through interrogation of the published literature. There has been a significant increase in publications in this area since the global pandemic. This will be discussed further in Chapter 5 of the thesis, when considering the feasibility of the exam going forward.

A pilot study is an opportunity to refine training and test training strategies (Polit and Beck 2017). In the pilot exam, the Exam Working Group was trained in Question-writing and standard setting and with the initial support and supervision of an educationalist, went on to deliver training to a wider group in preparation for the first EEMM. The value of conducting feasibility and pilot studies for implementation trials was explored by Pearson *et al* (Pearson, Naylor et al. 2020). They highlighted the importance of an iterative process, which is relevant to our feasibility study too. Feedback from candidates on exam standard was very positive, supporting our finding of high content validity and indicated that the exam blueprint was well aligned to the European Curriculum. Analysing the procedures and process identified areas which could be developed in order to enhance the quality of the candidates' exam experience. In response to candidate feedback, efforts were made to improve the image quality and improve the exam advertisement in the real exam. In an iterative process, candidates sitting the first real exam filled out a similar candidate experience questionnaire, which demonstrated more satisfaction with the exam content and image quality. Candidates' feedback from the pilot examination was an appreciated and invaluable component of the quality review process, integral to the assessment of feasibility.

A pilot study assesses the inclusion and exclusion criteria of the participants (In 2017). This was important in the pilot examination which was designed to be an 'exit' exam. Data looking at participant performance, assessed by number of years of training was

interesting, with some unexpected findings; overall, it showed a very weak relationship between candidate ability and training duration (Figure 4.13). When results were reviewed dividing candidates into trainees and consultants, the consultant group performed less well. There may be several reasons for this. It may be that the trainees are very focused on study and training progression, with access to regular training activities and therefore were more prepared through study. It is also possible that, with the change in some areas of the scope of practice, as demonstrated in Table 4.9, Table 4.10 and Table 4.11, there is harmonisation of the move away from the traditional 'bench-based' practice of CM towards a more clinically-centred practice, with ward rounds and MDTs.

When planning a pilot study, an appropriate sample size needs to be determined (In 2017). This was also important for the pilot exam to test the standard setting methods and achieve robust psychometric analysis to assess the validity and reliability of the exam (Nunnally and Bernstein 1994). Thabane *et al* have developed a number of tools to guide and support researchers when embarking on a feasibility study or pilot (Thabane, Ma et al. 2010). Such tools do not exist for undertaking a pilot exam as a feasibility study. I have reviewed the tools from Thabane and others (Thabane, Ma et al. 2010) (Skivington, Matthews et al. 2021) and adapted them to create a tool for groups developing a pilot exam (Table 4.8). I believe it provides a very useful overview of the complex factors required to make delivery of a quality assured high-stakes exam feasible. One of the objectives in developing the exam is the goal of harmonising training and practice in Europe.

Harmonisation of Training and Scope of Practice

The publication of the European Council Directive in 1993, Council Directive 93/16/EEC facilitates the free movement of doctors in Europe and the mutual recognition of their diplomas, certificates and other evidence of formal qualifications (EU 1993). The development of a legal framework in 2005, Directive 2005/36/EC was designed to allow EU citizens to transfer their qualifications and skills between member states. It works on the principle that a qualified professional in one member state is qualified to exercise the same profession in another state, and is based on the length of training in the individual Specialty and the title of qualification awarded at the end of training (Kortese 2016).

The European Union of Medical Specialists' (UEMS: Union Européenne des Médecins Spécialistes) vision is the harmonisation and improvement of the quality of medical specialists' training, qualifications and practice in the European Union. In response to this goal, the UEMS sections are developing European Training Requirements, European Curricula and European Exams. One of the challenges recognised by sections is the difference in training and/or scope of practice in specialist areas in Europe. A study of training in Child and Adolescent Psychiatry (CAP) in Europe, found that the structure of training was very different in the 34 countries surveyed. In several countries there was no structured training programme. CAP was not a recognised separate speciality in 34% of the countries (Karabekiroglu, Doğangün et al. 2006). The authors recognised the importance of quality teaching, a structured training programme and evidence-based evaluation of training to progress the process of harmonisation (Karabekiroglu, Doğangün et al. 2006). Specialists in radiation oncology also acknowledged the importance of their European core curriculum in harmonisation of practice in their specialist area (Baumann, Leer et al. 2004). The UEMS section of Public Health medicine outlined their strategy for

harmonisation of training and practice in Europe, starting with defining and harmonising the list of competencies needed to be a specialist in their field and to use the list to direct quality assured training (Westerling 2009). In Clinical Neurology, harmonisation is seen as political initiative, the medical care a patient can expect to receive will be of a similar standard in any country throughout the European Union (Cole and Kamondi 2023). To support the harmonisation journey, they recommend that national examinations assess the content of specified training modules (Cole and Kamondi 2023), rather than create a European exam.

In Clinical Microbiology, the role of the specialist can be divided into four key areas: Laboratory diagnostics and Management, Clinical Liaison and Antimicrobial Stewardship, Infection Prevention and Control, and Training and Education (Scott 2000, Raoult, Fournier et al. 2004, Doyle, Boyle et al. 2021).

The role of the CM has been evolving over the last four decades and has moved along a continuum from a largely laboratory-based role, with practical 'bench duties', a part of the routine work towards a more clinical role, bridging 'the gap' between the laboratory and the clinician, interpreting the Microbiology result to give clinical, diagnostic and Infection Prevention and Control advice (Mehtar 1995, Riordan, Cartwright et al. 2002). The rapid increase in novel, complex antimicrobial resistance mechanisms and the emergence of new pathogens has highlighted the importance of the CM role in giving advice regarding appropriate antimicrobial therapy, overseeing infection surveillance and management in the hospital and community and focusing on development, management and quality in the diagnostic laboratory (Humphreys, Nagy et al. 2010). The progression along the continuum from laboratory-based to a more clinical liaison role is different in different countries in the EU, prompting recognition of a need to develop professional

standards for Europe and harmonise training and scope of practice in Europe (Humphreys, Nagy et al. 2010, Doyle, Boyle et al. 2021).

UEMS section MM have taken a number of steps to promote and facilitate harmonisation of training and practice in MM in Europe. The European Training requirements for MM were completed in 2013 (UEMS 2013) and the European Curriculum in MM was ratified by council in 2017 (UEMS 2017), an important phase in harmonizing medical training (van der Aa, Scheele et al. 2019). A working group was established in 2017 to develop a European exam as a collaboration between UEMS and ESCMID.

The Medical Research Council (MRC) in the UK, recently published a new framework for developing and evaluating complex interventions (Skivington, Matthews et al. 2021). The MRC guidelines suggest that preliminary studies, including pilots, should be used to evaluate a package of interventions (such as an educational course), (Arain, Campbell et al. 2010). In Table 4.6 we explored our experience of the pilot exam looking at the MRC six phases of a complex intervention and the core elements of the intervention, Table 4.7. The MRC tools also provided a useful, assessment of the pilot exam and may be a useful starting point for future high-stakes exam pilots. The tool I have proposed as a detailed 'checklist' for groups setting up a high-stakes exam incorporates elements of the MRC in the feasibility criteria at the end of the tool. When applied to the educational intervention, the pilot exam, to assess the feasibility criteria, the feasibility of the intervention (Skivington, Matthews et al. 2021), the group should consider the optimal exam content, the exam platform, the acceptability of the exam to stakeholders, the likelihood of cost-effectiveness and the capacity of the Exam Working Group or committee to deliver the intervention (Table 4.8).

The MRC emphasizes the importance of the evaluation of the intervention, did the intervention work? Did it achieve its intended outcomes? (Skivington, Matthews et al. 2021). The objectives and outcome measures should be considered when developing a pilot exam. This is included in my tool for developing a high-stakes exam (Table 4.8). The outcome measures used in our exam were assessments of reliability, validity, psychometrics (Chapter 3) and surveys of candidate feedback, and scope of training and practice. Process evaluation, which considers fidelity and quality, can be evaluated using my proposed tool (Table 4.8). A complex intervention considers system level outcome measures, how a change in one part of the system affects behaviour in another, this could be changing and harmonisation of CM training and scope of practice.

Implementation should be considered early in the project. How the intervention, the exam, will be implemented in the 'real world'. This should be considered throughout all the steps; development, feasibility testing, process and outcome evaluation. In the exam group, we recognised the need to consider contextual factors that may hinder, e.g. cost to candidates, cost to the section of MM. We also recognize, flexibility is important in long-term intervention (Pfadenhauer, Gerhardus et al. 2017). It is likely that the exam will deliver this flexibility and concerns re cost and accessibility of the exam by alternating between an online exam and a face-to-face exam.

The new MRC framework identifies four perspectives; efficacy, effectiveness, theory based and systems, which can guide the how an intervention works in a complex system. A complex intervention such as the introduction of the European Exam in MM, can be considered as an event in a system. Systems are adaptive and what is important is how the system and intervention adapt to one another. How does the intervention, the exam, bring

about change? The efficacy of this intervention is simple: *'Assessment drives learning'*. The European Exam is aligned to the European Curriculum.

The scope of practice of Clinical Microbiology is a complex system and is adaptive. The field of training in Clinical Microbiology in Europe is part of the same complex system and is an adaptive system, this was demonstrated through the survey of training in CM in Europe, (Figure 4.5, Figure 4.6 and Figure 4.7), showing a move towards harmonisation of training in Europe (Doyle, Boyle et al. 2021), aligning to the European Training Requirements and European Curriculum.

Is the intervention, the exam, effective? The surveys of the scope of practice of Clinical Microbiology, show that clinical microbiology is adapting, that there is a move away from the traditional Laboratory-based practice of benchwork, towards a stronger base in clinical liaison. It is likely that one change is promoting the other (Skivington, Matthews et al. 2021). This is supported by the review of Country D, whose candidates did not perform as well as those from other countries in the exam (Figure 4.9).

Overall, the Consultants did not perform as well as trainees in the exam (Figure 4.10). My earlier suggestion was that this was due to the study and preparation undertaken by trainees. Another possibility is that the trainees' scope of training and therefore, practice is more aligned to the European curriculum, than the consultants who trained at a different time. In Country D, the survey on training provides evidence of a structured training programme, providing training opportunities in all the curriculum areas, but more laboratory-based, than clinically based. The scope of practice survey results for Country D suggests that it is moving along the continuum away from a laboratory benchwork-based, traditional Clinical Microbiology practice.

Harmonisation aims to provide standards for training and practice, while respecting autonomy (van der Aa, Scheele et al. 2019). Harmonising postgraduate medical education and training through common learning outcomes, in an outcome-based curriculum with systems for assessment, should occur while remaining cognisant of the context of training, the workplace. As a result, tensions may exist between standardisation and contextualisation in the journey towards harmonisation (van der Aa, Scheele et al. 2019). Some countries may not have the same training resources available as others, resulting in inconsistencies between the reality of training and the educational principles, e.g. in Obstetrics and Gynaecology training, access to training tools, such as simulation training may be different in different countries. While the authors do not propose to offer a solution to this challenge, they believe that understanding it may improve feasibility and sustainability of the harmonisation process (van der Aa, Scheele et al. 2019). Recognition of the tension that exists between standardisation and contextualisation in the development phase of the curriculum and recognising the need for flexibility implementation phase, may be help more countries progress on the journey towards harmonisation.

The European legislation to facilitate movement between EU countries has caused migratory tensions, with concerns for migration resulting in the 'brain drain' from the affected countries (García-Pérez, Amaya et al. 2007). A comprehensive paper describing the International migration and movement of doctor to and within OCED countries, 2000-2018, shows that migration is not the main cause of the global health workforce shortage (Socha and Dumont 2021). Ireland has one of the largest outflows of home-trained doctors, i.e. more than one doctor per 1000 population, along with Romania, Estonia and Iceland. This is similar to an earlier study, looking at Ireland in 2007, citing Ireland as the

European country with the highest percentage of doctors practising abroad (García-Pérez, Amaya et al. 2007), however, a recent publication tracking the Irish doctors leaving and returning to the Irish health system between 2015 and 2021, shows that the majority of them return (Pierse, Morris et al. 2023). While international migration may be worsening the crisis in some countries, reducing migration will not be the solution (Socha and Dumont 2021).

Not everyone favours harmonisation of medical practice in Europe as a goal. There is a view that harmonisation is not desirable and diversity is needed in healthcare, to generate innovation (Brearley 1995). However, the author also acknowledges that with harmonisation, there is an opportunity for the best features of one country's training systems to be shared and imported into those in other countries, but believes this is unlikely. The evidence is that harmonisation is happening and is improving the training opportunities and putting more focus on standards in training and meeting training needs. Developing international standards in healthcare is not new and implementing high standards in healthcare in all countries is desirable. The ultimate goal is well-trained doctors delivering safe, quality, evidence-based patient-centred care, while experiencing a dynamic, and professionally fulfilling career.

The creation of the European community of trainees that arises from the work of the ESCMID, the Trainee Association of ESCMID (TAE), is a platform for networking, sharing, comparing and innovation. The TAE, in a recent survey, highlighted the importance of a common plan and curriculum to harmonise training and improve quality of training (Palacios-Baena, Zapf et al. 2018). Trainees support the development of a European Exam (Yusuf, Ong et al. 2017) and believe that it will facilitate movement between countries in Europe (Yusuf, Ong et al. 2017).

Following on from the experience and evaluation of the pilot exam which support the feasibility of a real pan-European Exam in MM, as Allerberger references in his editorial, we have evidence of '*a success that speaks optimistically to a future of standardised, harmonised quality of training across Europe*' (Allerberger 2021). Harmonisation of training will support harmonisation of practice.

I believe that the pilot examination, while a feasibility study, does meet the criteria for a pilot study, with its focus on methodology and rigor of process. I also believe the proposed tool to assess feasibility in a pilot exam will be useful in guiding other exam development groups in this journey. Or perhaps, the tool proposed tool in this study, will join the ranks of papers described by Moore *et al* as "*problematic or inappropriate application of the term "pilot" to describe the study"* (Moore, Carter et al. 2011).

5. Discussion/Conclusion

5.1 Introduction

> *"Grab students by the tests and their hearts and minds will follow"* (Swanson and Case 1997).

Underpinned by the UEMS' vision of promoting free movement of medical specialists through the harmonisation of the highest level of specialist training and medical care, in 2017, the UEMS section of MM began the process of developing an exam. A pilot European Exam in MM was delivered online in March 2021, to assess the feasibility of creating a quality assured bench-mark pan-European assessment in MM.

There is an increasing focus on accountability and quality in healthcare and this is also true of medical education, seeing a welcome move away from opinion-based to evidence-based education (Boet, Sharma et al. 2012). With globalisation and interest in mobility of doctors and their training and interest among other stakeholders, such as patients and employers in medical education, a high-stakes exam should be legally defensible (Norcini, Anderson et al. 2018). An evidence-based education approach was applied to every step in the process of creating this high-stakes assessment. The approach to the curriculum design, the blueprinting, the question-writing, standard setting, psychometrics and results-feedback, were grounded in best practice and based on evidence from the rich bank of medical education research and standards in the literature.

Educational research should have the same process, protocols and rigor of clinical research (Boet, Sharma et al. 2012). The Kirkpatrick framework is a conceptual framework, first described in 1959, that can be used to categorise the impact of an educational

intervention (Kirkpatrick 1959). Barr *et al* modified it creating a six-level framework (Barr, Koppel et al. 2006). The second and third level explore the effect of the educational intervention on attitude, knowledge and skills. The fifth level considers the effect on the trainee's professional practice and the sixth level, the effect on the patient's condition or cost-effectiveness of the intervention. This modification reflects the shift away from a focus on how an educational intervention affects the learners satisfaction, to a focus on healthcare process and outcomes, required by stakeholders (Boet, Sharma et al. 2012). The educational intervention described in this paper, a pilot examination to explore feasibility of a high-stakes pan European Exam has impact at the highest levels of the modified Kirkpatrick framework, the learner's professional practice and delivery of patient care.

5.2 Summary of the Key findings

5.2.1 Feasibility

Robust standards and quality in assessment are essential components of high-stakes examinations. The serious consequences that can result from a high-stakes exam heighten the scrutiny. At the Ottawa Conference in 2010, a consensus was reached regarding seven criteria for good assessment that are applicable both to a single assessment or a system of assessment with one purpose (Norcini, Anderson et al. 2011). The criteria can be considered in the context of the pilot exam in MM. We demonstrated *validity*, in construct, content, face and process, which supports the use of the results (Chapter 3). We demonstrated *reproducibility* or consistency, supported by Cronbach's alpha (Chapter 3). Based on the rigorous process of development and supported by the test psychometric

analysis, we expect *equivalence*, that the same assessment will achieve equivalent scores across different cycles of testing. The exam *feasibility* has been demonstrated by the process of the pilot exam, a feasibility study, demonstrating that it is practical and realistic in the current context. An *educational effect* is anticipated. The trainees who participated in the pilot exam received detailed feedback which they can then use to focus their study activities for the EEMM. It should have a *catalytic effect*, driving future learning forward, in all nine areas of the curriculum. Finally, by keeping stakeholders informed of the project progress through presentations and publications and by gaining support of the largest European society in infection medicine, ESCMID, the exam has achieved *acceptability* from many of the stakeholders (Norcini, Anderson et al. 2011).

The conclusion from the pilot study to assess the feasibility of creating a pan-European Assessment in MM, is that a European Exam in MM is feasible.

5.2.2 Harmonisation of Scope of Practice

One of the 'problems' outlined at the start of this project was the challenge of recruiting CMs in Ireland, now and in the future, in order to meet the demands outlined in the recent National workforce review and plan (HSE and NDTP 2023). While the EU directive 2005/36/EC (EUR-Lex 2005) facilitates the movement of Medical specialists within the EU, it does not address the differences in training and practice. Differences in training and practice will deter the stakeholders and healthcare providers from recruitment of specialists from some EU countries. The scope of training and practice may be such that the specialist in one country lacks some of the knowledge and skills to deliver some of the patient care and tasks required in the role in a country outside that in which they trained.

This paper demonstrates that since the development of a European Curriculum in 2017, and development of a quality-assured, Knowledge-based European Exam, closely aligned to the curriculum there is a move towards harmonisation of training and practice in Europe (Chapter 4). The assessment which is blueprinted against all nine areas of the Europe curriculum, is assessing areas of the curriculum which may not yet be part of scope of practice in all countries in the EU, however, learners wishing to pass the European Exam will be assessed on all areas of the curriculum. It is possible, therefore, that success in the European Exam will make candidates more attractive to the stakeholders when applying for jobs outside their country of training. This may help address the challenge of recruitment of CMs faced in Ireland and some other EU countries. I acknowledge, however, that this conclusion, while possible, it is influenced by my preconceived idea that a motivated adult will elect to study topics, in preparation for the European exam that they would not otherwise consider reading.

However, it is also possible that they will continue the learner tradition of 'leaving out' material or sections of the curriculum, hoping that good achievement in other areas of the curriculum will compensate for their study and learning gaps. Of course, in the context of recruitment, if the candidate has passed the European exam, it is still the responsibility of the employer to check for knowledge gaps during an interview process.

How a student studies and what they learn is influenced by how they will be assessed, so assessment can be used as a strategic tool to direct the student studies (Cilliers, Schuwirth et al. 2010). In the current EEMM assessment model, the examinee can pass the exam, even if they fail the assessment of one or more parts of the curriculum, provided their overall score achieves the passing standard. It will be important to study trends in the exam results and scores in each curriculum area to see if there are obvious study gaps. If

this seems likely, the exam committee could consider making it a requirement that the examinee achieves the passing standard across all areas of the curriculum.

A recent review of the future of training in CM and ID raised concerns around equality, diversity and inclusivity for trainees (Last, Power et al. 2021). Diversity (socio-economic, gender, ethnicity) in society in Europe is a reality, and should be reflected in the healthcare professionals that deliver our healthcare (Mulder, Wouters et al. 2023). As ESCMID members, we have a responsibility to support the mission of the ESCMID Parity Committee, which promotes geographic balance within ESCMID and its related educational activities.

Ethnicity-related differences, or bias, in clinical skills assessment has been described in medical schools internationally (Low, Pollack et al. 2019) (Alexander, Osman et al. 2012). Both unconscious and explicit bias are well recognised pitfalls of 'Clerkship grading', which has demonstrated bias across gender, race and ethnicity, a bias which may continue into postgraduate medical assessment (Dayal, O'Connor et al. 2017, Bowe, Bly et al. 2022). Mulder et al, in their paper exploring Medical education in the Netherlands and why the demographics of specialists differ from the demographics of their medical students, considers 'similarity bias' and 'cultural cloning' as potential causes (Mulder, Wouters et al. 2023). The MCQ test format in the European Exam, which is delivered and scored electronically means that candidates are 'anonymous' when they are scored, therefore eliminating any 'similarity bias' and 'cultural cloning' that might occur.

Another concern when creating an assessment that will be delivered to candidates from many different countries is language (Avenia-Tapper and Llosa 2015, Holland and Stevens 2020). The exam was delivered in English, which would not be the first language for many examinees. In the candidate experience survey, we asked candidates about language and

the feedback was reassuring (Chapter 4; Figure 4.1). The exam-working group and question-writing groups have representation from many EU countries to minimize potential language bias.

An important consideration with respect to equality, diversity and inclusivity for trainees is the cost of sitting the exam. A report by the Academy of Medical Royal Colleges, focusing on cost of examinations to trainees in the UK, recommended that consideration be given to the principles of equality, diversity and inclusion when developing models of examination (Ellis, Scrimgeour et al. 2021). For trainees planning to sit the EEMM, feasibility will likely be judged based on cost and convenience (Norcini, Anderson et al. 2018). It is important that the exam is accessible to trainees in resource-poor countries.

5.2.3 Harmonisation of Training in Europe

While there is a trend towards harmonisation of training in CM in Europe, in terms of implementation of the European Training Requirements and the European curriculum, there is less harmonisation of the trainee experience while in training. A survey conducted by the Trainee Association of ESCMID (TAE), reported that while most trainees in CM in Europe were satisfied with their training programme (Yusuf, Ong et al. 2017), there are still some noted differences in some areas of training e.g. trainee supervision (Palacios-Baena, Zapf et al. 2018) and trainee mentorship (Ong, Zapf et al. 2019). There are also reported differences in personal life and working conditions (Maraolo, Ong et al. 2017). The trainees also reported that the implementation of the curriculum, at each training site can vary and 'low-level' implementation is recognised to be a major reason for dissatisfaction among trainees (Yusuf, Ong et al. 2017), however, this study was conducted prior to ratification of the European curriculum in CM in 2017. In 2021, five years after the ratification of the

European Curriculum, a survey providing an update on the specialist training in Clinical Microbiology in Europe has demonstrated a trend towards harmonisation of training (Doyle, Boyle et al. 2021) and it is anticipated that the European Exam will drive closer alignment between the training curriculum and the actual training situation.

5.2.4 A tool to develop and evaluate a pilot exam standard

The feasibility study was important to test the method of developing a high-stakes exam, grounded in evidence-base, quality assurance and standards. The 'tools' used to support researchers conducting and reporting a pilot study, a feasibility study (Thabane, Ma et al. 2010) and the framework for developing and evaluating a complex intervention (Skivington, Matthews et al. 2021) were useful in reviewing and evaluating the pilot exam, when modified to fit the model of an exam.

I have a proposed a resource (Chapter 4), that is based on modifications of the work of Thabane *et al* and Skivington et al, that can be used to guide and evaluate the process of developing a high-stakes exam (Thabane, Ma et al. 2010, Skivington, Matthews et al. 2021). This resource will be useful to other Clinician-Researchers undertaking this task in the future.

5.3 Limitations

> "*Surveys are perhaps the most used, and sometimes misused, methodological tools among academic researchers*" (Desselle 2005).

Survey tools were used as outcome measures in the project and four potential survey errors are well recognised; coverage errors, sampling error, non-response error and

measurement error (Harrison and Draugalis 1997) and "*Like the Apocalyptic four horsemen, you should strive to avoid all of them*" (Harrison and Draugalis 1997). The presence and extent of these errors will impose limitations on the data collected.

Sampling error was minimised by sending the survey to representatives from all twenty-seven countries that are eligible to sit the European Exam. Response rates to email questionnaires are often as low as 25%-30% but may improve with follow-up e-mails (Yun and Trumbo 2000). A recent meta-analysis of response rates in online surveys in the education-related field, reported an average response rate of 44.1% (Wu, Zhao et al. 2022). They do, however acknowledge that this is based on publications and therefore may result in an over-estimation of typical response rates.

The candidate experience survey, conducted as part of the feasibility study in the pilot exam had 100% response rate, simple questions and provided useful feedback for the exam pilot project, without any non-response bias (see Chapter 4). The survey exploring specialist training in CM in Europe had responses from 70% of the target countries. Multiple contacts and personalisation were used to improve the response rate (Dillman 2000). However, the resulting non-response bias of 30% may affect the reliability and validity of the data. The scope of practice surveys had a response from 59% (16/27) of countries surveyed in 2019 and 2022 respectively, resulting in a potentially high non-response bias. The initial invitation to participate was followed by multiple reminders, but there were still countries for which no data was submitted. The scope of practice survey was quite long, with 69 questions, which may have had a negative relationship with the response rate (Burgard, Bošnjak et al. 2020).

Non-response bias is recognised as a key limitation in surveys as a research tool (Burmeister 2003). The countries that responded were compared to the countries who are

members of the UEMS section of MM. The countries were divided into five areas, using the ESCMID description of Geographic Regions (Commission 2012). When divided into five geographic areas, the countries that responded in both the scope of practice surveys and the training survey aligned well with the membership of the section, so the nonresponse bias may be less than expected (see Chapter 4, Figures 4.4 and 4.8).

Other limitations in the scope of practice surveys include; the survey content, how questions were phrased and the use of different terminology to describe practice or staff roles in different parts of Europe e.g. the term laboratory scientist versus laboratory technician (Belson 1981). It may have been prevented by providing the respondents of a clear definition of the terms with the question. This can reduce respondent misinterpretation (Peytchev, Conrad et al. 2010). Although the surveys had limitations, it is recognised that survey research can be very powerful and may be the only way to conduct a body of research (Draugalis, Coons et al. 2008).

In my role as the researcher I am considered an 'insider', which may affect my ability to interpret the data in a neutral manner and may therefore be considered a limitation of this work (Boet, Sharma et al. 2012). It is important therefore, that I engage in reflexivity, to consider how my subjectivity, or bias, and contextual factors might shape my inquiry (Olmos-Vega, Stalmeijer et al. 2023). My perspectives are provided by my professional and personal contexts; my desire for the option of an alternative exam and the recruitment challenges in Ireland, and these may shape the work. Adopting a reflexive approach to the project, monitoring my research processes and getting feedback from those whose opinion I value and considering their reaction to the work, are some of the steps I have taken to keep my report as free as possible from researcher bias.

5.4 Recommendations

5.4.1 Next exam online with remote invigilation

Assessment is an iterative process. Learning from the experience of the pilot exam, there are a number of considerations for future diets of the European Exam in MM. The feasibility analysis determined a EEMM to be feasible with modifications. The ADePT analysis explored these modifications, highlighting concerns around remote invigilation and security of the exam content (Chapter 4, Table 4.4). The immediate solution was to deliver the first EEMM in a face-to-face setting, so the exam was delivered in Paris in May 2022.

The main concern with delivering an online test is preserving the academic integrity of the exam. Academic honesty in higher education is not a new concern. The six most commonly used cheating practices, reported from a Norwegian survey were found to be: impersonation, forbidden aids, peeking, peer collaboration, outside assistance and student-staff collusion (Chirumamilla, Sindre et al. 2020). Some countries, e.g. Australia, had explored delivery of exams online with remote invigilation prior to the COVID-19 pandemic (Reedy, Pfitzner et al. 2021). During the pandemic, many institutions moved their exams to an online remote E-exam format (Choi, Jegatheeswaran et al. 2020, Papapanou, Routsi et al. 2022).

Online proctoring was developed to reduce the risk of academic dishonesty in E-exams (Arnò, Galassi et al. 2021, Fawns and Schaepkens 2022). A proctoring system is often a person or group of people monitoring the student remotely, however, some commercial proctoring systems also rely on algorithms to supervise students, an automated system which creates an alert if it detects unusual or suspicious behaviour, e.g. the student leaving

the room or having a conversation or may even use an algorithm of pattern recognition, monitoring 'clicking behaviour' or disconnection from the internet (Andreou, Peters et al. 2021, Arnò, Galassi et al. 2021). A study comparing exam scores from proctored and un-proctored online exams, in the same student group across two semesters, found that there were significant differences between the exam scores from the same class, and concluded that proctoring was necessary to prevent dishonesty (Reisenwitz 2020). Academic integrity and security are essential in a high-stakes exam. In some universities and institutions, proctoring of online exams may be required to meet the requirements of the programme accreditation bodies (Kinley 2023).

A study in Belgium during the Flemish Advanced Master's in General Practice directly compared a remote versus an onsite proctored exam. The remote proctoring system recorded the microphone, computer screen and camera. The proctoring software could analyse behaviours, using an algorithm of pattern recognition to monitor the candidate responses and could facilitate live proctoring. The human proctor could join the candidate by live feed if they had any concerns about their behaviour (Andreou, Peters et al. 2021). In this large study of 593 candidates, the exam results of the remote proctored group were not significantly different to the on-site proctored group. Jaap *et al* reported similar findings in Edinburgh when they compared remote versus onsite invigilation of their summative assessment (Jaap, Dewar et al. 2021).

There were a number of difficulties described with the online remote E-exam, particularly technical problems relating to internet connectivity (Reedy, Pfitzner et al. 2021). Candidates in both the Belgian and Scottish studies were surveyed to assess if the remote-proctoring impacted on their emotional well-being. Some candidates reported anxiety relating to potential technical issues, some reported anxiety around the feeling of 'being

watched', but other emotions were positive, with candidates being reassured by the proctoring software (Andreou, Peters et al. 2021, Jaap, Dewar et al. 2021). Other authors observed similar themes among students sitting exams with remote proctoring, with some examinees reporting a negative educational experience (Milone, Cortese et al. 2017, Elsalem, Al-Azzam et al. 2021). If the online proctoring system involves recording the student, there may be concerns regarding privacy and data protection or there may be personal and cultural reasons why students may not wish to be watched (Guangul, Suhail et al. 2020).

In the pilot exam, in the anonymous post-exam candidate feedback questionnaire, the majority of candidates reported that the remote proctoring seemed secure and that it would be difficult to cheat (Chapter 4, Figure 4.2), so despite the exam project group concerns, this was somewhat reassuring. The findings of the Belgium high-stakes postgraduate professional exam study are also reassuring (Andreou, Peters et al. 2021). The plan going forward is to alternate between online remotely invigilated exams and face-to-face exams. The use of a robust remote-proctoring system with a plan for remedial actions should a problem occur e.g. candidates experiencing technical issues might be offered extra time (Jaap, Dewar et al. 2021), will address some of the concerns in the project group regarding academic misconduct and among examinees regarding technical issues. An online remotely proctored exam would make the exam more accessible to candidates from resource-poor countries, by eliminating the need for travel and accommodation, supporting the principles of equality, diversity and inclusion and if a trainee is concerned about technical problems or privacy, there is still the option of sitting the exam, in person, on alternating diets of the exam.

5.4.2 Expand exam to include other test formats.

A single method of assessment is unable to assess all the attributes required to become a competent health professional (Cox, Irby et al. 2007, Norcini, Anderson et al. 2018). Currently, the EEMM is a single objective test, testing Knowledge learning outcomes at the first and second level of Miller's pyramid (Miller 1990). A systems-based framework has been suggested to blend single assessments in order to achieve the most benefit for all stakeholders (Norcini et al., 2018). It is recommended that the CM trainee undergoes formative assessment, to assess their skills and attitudes during the training programme. The survey of training reported that 49-67% of trainees had formative assessments during training and all trainees kept a logbook or checklist of training activities. However, the EEMM assessors do not specifically assess trainees' compliance with this recommendation, so it is possible that aspects of the curriculum are not currently assessed.

In a survey undertaken by CESMA in 2021, 20 out of 34 UEMS exams had two or more parts to the assessment (Durkan 2023). The FRCPath UK exam has two parts, Part 1 is an MCQ, delivered online with remote invigilation while Part 2 includes an OSCE, delivered face-to-face and is paper based. During the Autumn 2020 sitting of the FRCPath Part 2, delivery was online, with remote invigilation (RCPath(UK) 2020).

Once the EEMM MCQ exam is established, I believe a second part to the exam e.g. an Objective Structured Practical Exam (OSPE), which could include stations on communication, would add significantly to the robustness of the EEMM (Malhotra, Shah et al. 2013). This would allow the assessment blueprint to include knowledge (at a higher level of Miller's pyramid (Miller 1990), skills and behaviours, creating a system of assessment that is constructively aligned to the entire curriculum (Biggs 2014). The

assessment, which is designed to be end/near end of programme, will be calibrated to the endpoint standard and learning outcomes (Norcini, Anderson et al. 2018). The system of assessment will also continue to include a workplace-based portfolio or log-book, to improve the quality and consistency of educational standards and training (Schüttpelz-Brauns, Narciss et al. 2016).

The development of a system of assessment, with purposeful blueprinting, driven by the desired programme outcomes in knowledge, skills and attitude may also provide a solution to the second problem outlined in my introduction, and may offer the Specialist training programme in Clinical Microbiology in Ireland the choice of an alternative exam to the UK exam.

5.5 Future research/work

5.5.1 Scope of practice update

The two scope of practice surveys not only demonstrated a trend towards harmonisation of scope of practice but were also provided a very useful insight into some of the results in the pilot exam. At the start of the project, consideration was given to having two versions of the exam, with different weightings in each blueprint, to facilitate countries where scope of practice, excludes an entire curriculum area, e.g. Infection Prevention and Control. Other sections of UEMS have also considered this when planning their exam (Cole, Kamondi et al. 2022, Cole and Kamondi 2023). The scope of practice survey showed that the divergence was not as great as initially anticipated and following consideration of what was 'fair' to examinees and the goal of harmonisation, it was agreed that there would be one blueprint, which would strike a balance.

The Scope of Practice survey will be repeated in 2025, after the next diet of the EEMM. The scope of practice survey may be used to revise the weighting in the blueprint of the next diet of the exam. Pre-contacting the potential respondents prior to the survey can yield a higher response-rate (Burgard, Bošnjak et al. 2020, Wu, Zhao et al. 2022), so this is something that we will aim to do. Another option would be to ask respondents to complete the survey while attending the section meeting in 2025, as face-to-face contact and personal contact can increase a survey response rate (Cook, Heath et al. 2000).

5.5.2 Curriculum revision

The European curriculum in Clinical Microbiology was ratified by UEMS council in October 2017. The curriculum review planned for 5 years post ratification was delayed due to the work of Clinical Microbiology in the global pandemic. The curriculum will be reviewed in 2024, to incorporate the rapidly evolving challenges and developments experienced in CM since its development. Trainees in CM have identified a number of areas that they feel require consideration in the review, including management and health economics (Yusuf, Ong et al. 2017), an update on antimicrobial resistance, One Health and Digital technologies (Last, Power et al. 2021). Other rapidly evolving areas include rapid diagnostics and genomics in CM. The curriculum update will be undertaken by a working group of the UEMS section of MM, working closely with the exam working group, Action research is an approach which tries to put evidence-based educational innovations into practice, to improve teaching, often in cycles. Action research has been used to improve curricula at all levels of education, from elementary school to university (Laudonia, Mamlok-Naaman et al. 2018). The action research approach may be useful for this review, using the checklists described by Zuber-Skerritt and Fletcher, for quality action research (Zuber-Skerritt and Fletcher 2007), aiming to produce an outcome-based framework suitable for creating the blueprint for the system of assessment, continuing the journey towards harmonisation of quality training and practice in Europe.

5.6 Researcher perspective and learning

The UEMS section MM, in a collaboration the European Society of Clinical Microbiology and Infectious Diseases (ESCMID), held the first European Exam in MM in Paris on the 23rd of May 2022, the final outcome measure in the feasibility study.

5.6.1. Quality Assurance is key, a social imperative.

When we consider validity in assessment in a more holistic way, we consider the consequences for not just the learner but also the impact at the highest level of Kilpatrick's framework, the patient and society (Marceau, Gallagher et al. 2018).

The measuring and monitoring of healthcare professionals competencies is increasingly becoming of interest to society (Dijkstra, Van der Vleuten et al. 2010). It is Important to ensure that healthcare professional high-stakes assessments which lead to certification are quality assured. Society is concerned about patient safety and the validity of evidence, so the evidence used to demonstrate quality in the assessment should be credible by society. Validation should be embedded in the entire process of assessment, from blueprinting to question selection and interpretation of scores (Dijkstra, Van der Vleuten et al. 2010).

We must also consider the consequential validity, which may be beneficial or harmful, intended or not intended. An error in a high-stakes assessment may result in a learner being licensed to practice who has not yet achieved mastery in the necessary competencies. Philips modification of Kirkpatrick's framework considers the stakeholders perspective, (Patient, Public, employers, UEMS, ESCMID, Trainees etc), with the fifth level being 'return on investment' (Phillips 2012). There is an increasing appetite for

transparency and interest among stakeholders such as employers and the public in the exam process (Norcini, Anderson et al. 2018).

Kaufman and Keller's modification of Kirkpatrick's framework, suggests that the fifth level of evaluation should be '*societal value* '(Kaufman and Keller 1994), supporting a growing feeling that validity in healthcare professionals' assessment is a social imperative.

5.6.2 Work may help other clinician researchers developing a high-stakes exam.

The study of feasibility through the pilot exam was an important test of the process involved in developing a quality exam. The pilot exam was also a journey of learning for the exam project team who are clinician-researchers. Many other high-stakes UEMS examinations are developed and delivered by clinician-researchers. I believe that the pilot exam tool, developed during this work will support and guide colleagues to work to a high standard and create an exam that is robust, defensible, quality-assured and feasible.

Appendices

Appendix 1: Candidate Survey

Exam process

1. How did you hear about the EEMM?

- [ ] ESCMID
- [ ] Microbiology society
- [ ] UEMS website
- [ ] Other (please specify)

2. How easy was the registration process?

- ◯ Very easy
- ◯ Somewhat easy
- ◯ Not at all easy

Comments?

3. Did you find the information regarding the exam clear?

- ◯ Very clear
- ◯ Somewhat clear
- ◯ Not at all clear

Comments

4. Did you find the exam delivery system easy to use?

- ◯ Agree
- ◯ Neither agree nor disagree
- ◯ Disagree

Comments

5. How do you rate the instructions and information on how to use the exam delivery system?

- ◯ Very clear
- ◯ Somewhat clear
- ◯ Not at all clear

Comments?

6. How do you rate the question layout/format?

- ◯ Very satisfied
- ◯ Satisfied
- ◯ Neither satisfied nor dissatisfied
- ◯ Dissatisfied
- ◯ Very dissatisfied

Comments?

7. How do you rate the quality of the images?

- ◯ Satisfied
- ◯ Neither satisfied nor dissatisfied
- ◯ Dissatisfied

Comments

8. What did you think of the language used in the questions?

- ◯ Extremely clear
- ◯ Somewhat clear
- ◯ Not at all clear

Comments

9. What do you think of the supervision/invigilation

- ◯ Very effective
- ◯ Somewhat effective
- ◯ Not at all effective

Comments?

10. Do you think it would be possible for a candidate to cheat during the exam?

- ◯ Definitely would
- ◯ Probably would
- ◯ Definitely would not

Comments?

11. Do you have any other comments or concerns regarding security of the exam or how it might be improved?

12. Do you have any comments or suggestions regarding the exam delivery system or how it might be improved?

13. Did you think the exam content was easy or difficult?

- ◯ Very easy
- ◯ Easy
- ◯ Neither easy nor difficult
- ◯ Difficult
- ◯ Very difficult

Comments?

14. This is an 'exit exam', intended for candidates in year 4-5 of training or Consultants. Do you think the exam is the right standard for this group?

- ◯ Very easy
- ◯ Easy
- ◯ Neither easy nor difficult
- ◯ Difficult
- ◯ Very difficult

Comments

15. Do you think the exam tested your knowledge of all aspects of Microbiology and the European Microbiology curriculum?

- ◯ Strongly agree
- ◯ Somewhat agree
- ◯ Somewhat disagree
- ◯ Strongly disagree

Comments?

16. The exam tested the knowledge needed to work as a Microbiology specialist?

- ◯ Strongly agree
- ◯ Somewhat agree
- ◯ Somewhat disagree
- ◯ Strongly disagree
- ◯ Comments?

There are 9 sections in the European Curriculum, were they all tested in the exam?

17. Knowledge of **General Microbiology** was tested adequately?

- ◯ Strongly agree
- ◯ Agree
- ◯ Disagree
- ◯ Strongly disagree

comments?

18. Knowledge of **Bacteriology** was tested adequately

- ◯ Strongly agree
- ◯ Agree
- ◯ Disagree
- ◯ Strongly disagree

Comments?

19. Knowledge of **Virology** was tested adequately

- ◯ Strongly agree
- ◯ Agree
- ◯ Disagree
- ◯ Strongly disagree

Comments

20. Knowledge of **Parasitology** was tested adequately

- ◯ Strongly agree
- ◯ Agree
- ◯ Disagree
- ◯ Strongly disagree

Comments

21. Knowledge of **Mycology** was tested adequately?

- ◯ Strongly agree
- ◯ Agree
- ◯ Disagree
- ◯ Strongly disagree

Comments

22. Knowledge of **Clinical Medicine** was tested adequately?

- ◯ Strongly agree
- ◯ Agree
- ◯ Neither agree nor disagree
- ◯ Disagree
- ◯ Strongly disagree

Comments

23. Knowledge of **Infection Prevention and Control** was tested adequately?

- ◯ Strongly agree
- ◯ Agree
- ◯ Disagree
- ◯ Strongly disagree

Comments?

24. Knowledge of **Antimicrobials** was tested adequately?

- ◯ Strongly agree
- ◯ Agree
- ◯ Disagree
- ◯ Strongly disagree
- ◯ Comments

25. Knowledge of **laboratory management** was tested adequately?

- ◯ Strongly agree
- ◯ Agree
- ◯ Disagree
- ◯ Strongly disagree

Comments

26. If you chose **disagree** or **strongly disagree** for any of the 9 curriculum areas, could you indicate how the assessment might be improved?

- ☐ More MCQs
- ☐ More images
- ☐ Written paper
- ☐ Oral examination/interview
- ☐ Addition of a logbook/record of training activities/experience

Other (please specify)

27. Does the exam test the knowledge required to work as a Clinical Microbiologist in your country?

- ◯ Yes it does test the knowledge required to work as a Clinical Microbiologist in my country
- ◯ No it does not test the knowledge required to work as a Clinical Microbiologist in my country
- ◯ Other (please specify)

General information

* 28. How many months specialist Microbiology training have you completed?

29. Did you work as a general/ward-based doctor before starting Microbiology specialist training?

- ◯ Yes
- ◯ No

If yes, for how many months?

* 30. Have you taken any other specialist professional Microbiology/infection exam in the last 6 months or intend to sit one in the next 12 months?

- ◯ Yes
- ◯ No

If yes, please specify the name of exam

* 31. How do you think the standard of this exam compares to the standard of the professional exam you have sat/will sit?

- ◯ Easier than the the professional exam I will sit/have sat
- ◯ The same difficulty as the professional exam I will sit/have sat
- ◯ More difficult than the professional exam I will sit/have sat

Comments?

* 32. What suggestions do you have for improving this exam?

- ☐ Advertisement?
- ☐ Delivery platform?
- ☐ Supervision?
- ☐ Content?

Other (please specify)

Appendix 2: Scope of Practice Survey

Introduction

1. I have received the Participant Information leaflet

- [ ] Yes
- [ ] No

2. I agree to participate in this survey and consent to the data collected being used and stored as outlined in the Participant Information leaflet.

- ◯ Yes
- ◯ No

3. What country are you from?

4. Please estimate how many Consultant Clinical Microbiology medical specialists there are in your country.

- [ ] 0-20
- [ ] 21-50
- [ ] 51-100
- [ ] 101-200
- [ ] 201-300
- [ ] 301-400
- [ ] 401-500
- [ ] >500

5. Is there a National training programme in clinical Microbiology in your country?

- ◯ Yes
- ◯ No

6. Is there a National or Regional lead who oversees training to become a medical specialist in Medical Microbiology?

	National	Regional
National/Regional		

If No, how is training to be a Medical Microbiologist structured?.

7. Do your trainees currently sit an exam before they can be considered a specialist in Medical Microbiology?

- ◯ Yes
- ◯ No

8. Does your country have a specialist Medical Microbiology exam that they sit?

- ◯ Yes
- ◯ No

9. Do your trainees sit a specialist Medical Microbiology exam run by another country?

- ◯ Yes
- ◯ No

10. How many years do your trainees work as a Clinical/General doctor before entering training to become a specialist in Medical Microbiology?

- ☐ 0 years
- ☐ 1 year
- ☐ 2 years
- ☐ 3 years
- ☐ 4 years
- ☐ 5 years

11. How many years do your trainees spend in specialist training in Medical Microbiology?

- ☐ 0 years
- ☐ 1 year
- ☐ 2 years
- ☐ 3 years
- ☐ 4 years
- ☐ 5 years
- ☐ 6 years

12. Are there any special requirements your trainee MUST complete BEFORE starting training to become a specialist in Medical Microbiology?

	Clinical medicine year(s)	Clinical medicine postgraduate exams e.g. MRCP	Masters degree	PHD
1				

Other? Please specify

13. Do all Medical Microbiology Laboratory results in your country have to be 'overseen' or 'certified' by a Medical Microbiologist?

- ◯ Yes
- ◯ No

14. Is it a Legal requirement in your country that all Medical Microbiology Laboratory results have to be 'overseen' or 'certified' by a Medical Microbiologist?

- ◯ Yes
- ◯ No

15. Are Microbiology Laboratorys in your country

- ◯ Funded and overseen by the state or health service
- ◯ Funded privately and run privately
- ◯ Other (please specify)

General /Laboratory

Please indicate below which person is most likely to carry out/be responsible for the roles/tasks described below in your country:

* 16. Specimen set up

	Never (or)%)	Infrequently (1-5%)	6-25%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	☐	☐	☐	☐	☐	☐	☐
Infectious diseases	☐	☐	☐	☐	☐	☐	☐
Laboratory Scientist	☐	☐	☐	☐	☐	☐	☐
Nurse	☐	☐	☐	☐	☐	☐	☐
Pharmacist	☐	☐	☐	☐	☐	☐	☐
Laboratory person (not scientist, not medical)	☐	☐	☐	☐	☐	☐	☐
Other	☐	☐	☐	☐	☐	☐	☐

Other (please specify)

* 17. Developing Standard Operating Procedures/protocols for specimen set up, culture conditions and microscopy.

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 18. Performs the Microscopy

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 19. Performs the Molecular assays

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 20. Performs the antimicrobial susceptibility tests

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 21. Decides when susceptibility testing is required

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 22. Decides when extra sensitivity testing is required

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 23. The Standard Operating Procedures/Protocols for antimicrobial sensitivity test choices are decided by

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 24. The Standard Operating Procedures/Protocols for Identification of organisms are decided by

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 25. Performs the tests for identification of organisms e.g. set up API or MALDI-Tof

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 26. Develops the Standard Operating Procedures/Protocols for serology assays

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 27. Performs the serology assay/test e.g. for HBsAg

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 28. Performs the examination of blood/stool other tissue for the presence of protozoa/helminths

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 29. Developing Standard Operating Procedures/Protocols for the examination of blood/stool/other tissue for the presence of protozoa/helminths

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 30. Performs the tests for the examination of skin/hair/nails

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 31. Reports and approves the FINAL Microbiology result/report

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 32. Interpret the test results to give clinical advice

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 33. Interpret the test result to give infection prevention and control advice

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 34. Decides when a sample/organism will be referred to a reference facility for additional testing

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

Antimicrobials

Please indicated below which person is most likely to carry out/be responsible for the roles/tasks described below in your country:

* 35. Developing guidelines/protocols for Surgical prophylaxis

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 36. Developing guidelines/protocols for empiric antimicrobial choices

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 37. Oversees surveillance of antimicrobial resistance

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

Infection Prevention and Control in Hospital and Community

Please indicate below which person is most likely to carry out/be responsible for the roles/tasks described below in your country:

* 38. Lead role in Hospital outbreaks of infection

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 39. Lead role in Community outbreaks of Infection

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 40. Lead role in matters relating to hospital building and design

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 41. Lead role in matters relating to Hospital ventilation requirements e.g Standards for Ventilation in the Operating Theatre

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 42. Leads the Infection Prevention and Control team in the Hospital

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 43. Leads the Infection Prevention and Control team in the Community

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 44. Gives advice on patient isolation requirements

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 45. Gives advice on prevention of device-related infection

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 46. Leads on development of Guidelines/Protocols for Infection Prevention and Control in the Hospital/Community

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 47. Leads on development of Guidelines/Protocols for preventing transmission and Nosocomial spread in the Hospital.

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 48. Leads on development of Guidelines/Protocols for decontamination of equipment in the hospital

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 49. Review and implementation of new or revised National and International Guidelines in Infection Prevention and Control e.g. Guidelines for Management of CPE

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

Clinical Medicine

Please indicate below which person is most likely to carry out/be responsible for the roles/tasks described below in your country:

* 50. Takes a clinical/infection history from the patient or patients primary clinician regarding the patient presenting with an infection/infection syndrome

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 51. Carries out a physical examination of a patient presenting with an infection/infection syndrome.

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 52. Advises on the diagnosis of a patient with infection of a patient presenting with an infection/infection syndrome

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 53. Advises on the treatment of a patient presenting with an infection/infection syndrome

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 54. Advises on the prevention of infection e.g. vaccination

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 55. Participates in a ward round in the Critical Care Unit patients

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 56. Participates in a ward round of other specialist units e.g. transplant ward, burns unit

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 57. Participates in a ward round of non-critical care patients with a presumed infection/infection syndrome

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 58. Participates in multidisciplinary team meeting (MDT)/ward round with specialist services e.g. Haematology, Orthopaedics

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 59. Gives advice to clinicians by phone on diagnosis and management of infection

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 60. Gives advice to General Practitioners on diagnosis and management of infection

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 61. Gives advice on community-acquired and Nosocomial infections in which environmental factors play a role (e.g., food, water, air)

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

Laboratory Management

Please indicate below which person is most likely to carry out/be responsible for the roles/tasks described below in your country:

* 62. Lead role in management of the laboratory

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 63. Lead role in laboratory accreditation

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 64. Lead role in overseeing quality in the laboratory

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 65. Lead role in development and review of the laboratory Standard Operating Procedures/Protocols

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 66. Lead role in introduction of new technologies, platforms, tests in the laboratory

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

* 67. Lead role in ensuring safety in the laboratory

	Never (or 0%)	Infrequently (1-5%)	6-25%	26-50%	51-75%	76-95%	Yes, almost always 96-99%	Always (or 100%)
Medical Microbiology	○	○	○	○	○	○	○	○
Infectious diseases	○	○	○	○	○	○	○	○
Laboratory Scientist	○	○	○	○	○	○	○	○
Nurse	○	○	○	○	○	○	○	○
Pharmacist	○	○	○	○	○	○	○	○
Laboratory person (not scientist, not medical)	○	○	○	○	○	○	○	○
Other	○	○	○	○	○	○	○	○

Other (please specify)

68. Is Medical Microbiology taught as a 'stand alone' or separate module in medical schools in your country

- ○ Yes it is always a separate module
- ○ No it is never a separate module.
- ○ In some medical schools it is separate, in some it is not
- ○ Other (please specify) or add comment for clarification

69. Who is primarily responsible for delivering the Microbiology teaching in your Medical schools

- ○ Consultant Microbiologist
- ○ Microbiology trainees
- ○ Laboratory Scientists
- ○ Post-doctorate
- ○ Nurse
- ○ Other
- ○ Other (please specify) or add comment regarding your answers

Appendix 3: Exam Delivery Platform- Criteria for Selection

The following template was created by the treasurer of the working group with input from the other members

Exam delivery platform:

Background

The Medical Microbiology (MM) Section of UEMS wishes to launch an examination in MM.

UEMS is inviting examination assessment suppliers to provide detailed descriptions of the product(s) that you can provide with the associated costs and how you wish to work with UEMS Section of MM to deliver the examination. Companies are invited to complete the following questionnaire. We are proposing to run a pilot examination in April 2020.

Set up costs for this examination are being provided by ESCMID (European Society for Clinical Microbiology and Infectious Diseases) with the expectation that the examination will become self-funding within 4 years (i.e. from the April 23 Diet).

Companies are requested to describe how they would meet the requirements listed.

Assumptions presented to the companies:

The exam will be electronic, delivered on desktop-top, lap-top or tablet computers.
It shall be set in English only (in the first instance)
For the purposes of this proposal, assume that the examination shall comprise a single paper of 120 multiple choice questions (e.g. SBAs, EMQs), lasting 3 hours.
We anticipate that it will be attempted by up to 50 candidates in the first instance.
It will be delivered in one location annually at various European cities, depending on the location of the annual European scientific conference, ECCMID (European Congress of Clinical Microbiology and Infectious Diseases). The meeting is normally held in April. The location of forthcoming conferences can be checked on the ESCMID website. The April 2020 meeting will be held in Paris.
Sufficient rooms to enable the exam to be taken will be provided.
Invigilators will be provided (except as specified below).
Please assume that the examination will be held on the Friday morning, prior to commencement of ECCMID. The Exam Committee will meet at 17:00hrs on the Tuesday immediately after the close of ECCMID
The contract for the examination delivery will be with, and signed by, UEMS. Their requirements will need to be met.
The precise detail of the examination development process, to include question writing, initial and subsequent individual question review, initial and subsequent standard setting of individual questions, development of the whole examination, setting and reviewing the examination pass mark, etc has not been finalised. We are likely to utilise a statistical standard setting procedure, e.g. Angoff
For the purposes of this invitation, the examination will be run by an Examination Committee comprising 5 senior consultant microbiologists, referred to as ‘Exam Committee’ in this document. Further consultants may be co-opted from time to time. Questions will be provided by a network of ‘Question writers’ who will submit questions remotely.
The Companies submitting descriptions of their products are referred to as 'Commercial Partners' abbreviated to 'Partners'.

The product or products that Companies propose to provide will be referred to as 'Solutions'. This includes software, any associated advice and exam delivery.
The term 'Working Day' refers to Monday to Friday 09:00 to 17:00hrs local time.
Multiple choice question (MCQ) types: Single best answer (SBA), Extended matching questions (EMQs).

Proposal structure

The request for proposals is divided into sections which follow sequentially

- Question database
- Examination delivery (including call for examination entry)
- Post examination results analysis
- Issue of results

The Question database		Essential or Desirable
1	The database will have at least 2 levels of access: one level for question writers who will only be able to access questions that they have written; and a higher-level allowing access to all questions	E
2	Please state the maximum number of individual Question Writers and Exam Committee members that the database can accommodate individually	E
3	The database will be able to accept questions in the following formats: SBA (of 4 or 5) or EMQs (up to 10 choices to multiple stems).	E
4	Each question can comprise text, alphanumeric data in tabular format and / or images. Please describe any limitations including image size / resolution	E
5	Each question can be tagged by multiple identifiers to enable linkage with specific sections of the Curriculum	E

6	The name of the question writer and date/time submitted will be 'stamped' on each question and retained in the database.	E
7	The database will keep an audit trail of each occasion that a question has been viewed or amended (by whom, and when).	D
8	Questions within the database will be marked or flagged as: a) Draft (i.e. submitted by Question Writer), b) Screened (i.e. screened for potential suitability by a Board member) with Provisional Acceptance, Returned, Rejected and c) Reviewed (i.e. discussed by Question Review Group). Alternative terminology is acceptable.	E
9	The database will allow Draft questions to be returned to the author for amendment / clarification with a brief message from the screener.	D
10	Each question will have a field where various numerical and textual information can be stored (e.g. the author's proposed estimated level of question difficulty on a predetermined scale).	D
11	Questions can be selected for inclusion at a Question Writers Review meeting.	D
12	The date of discussion at the Question Writers meeting can be stored	D
13	Following Discussion at the Question Writers' review group updated ratings can be stored	D
14	Questions selected for potential inclusion in a particular diet of the examination can be flagged as such.	E
15	Once candidate questions are selected to populate a particular diet of the examination, the 'draft' examination can be made available for testing by independent individual(s) as nominated by the Board	E
16	The database must be secure at every stage of the question writing and examination development. Please describe your approach.	E

Logistics of Examination Delivery

Logistics of examination delivery		Essential or Desirable
1	The Partner shall receive applications from candidates wishing to sit the examinations	E
2	The Partner shall screen applications in accordance with criteria laid down by the Board	D
3	Candidate demographics (to include as a minimum, stage in training or specialist training completion, national association membership at time of application, nation of primary medical degree qualification, number of previous attempts) are assigned and stored to facilitate subsequent analysis.	D
4	The Partner shall refer applications that do not meet set acceptance criteria to a nominated person for a decision whether to accept	D
5	The Partner shall receive and process payments for candidates.	D
6	If the Partner receives payments on behalf of UEMS it shall remit all payments received, without any deductions, or as agreed. A Statement / Audit of finances to be presented quarterly.	E
7	The Partner shall undertake all communications with candidates with respect to examination logistics within an agreed timetable. This will include preparations for reasonable adjustments.	E
8	The Partner shall provide all hardware and software for all candidates within an agreed maximum capacity (50 for the purposes of this proposal)	E
9	The Partner shall provide a specification for the room(s) as required to run the examination	

10	The Partner shall take responsibility to ensure that the room(s) provided meet the Partner's requirements to run the examination effectively	E
11	The Partner shall deliver the examination on site. This will include candidate identification and registration.	E
12	The Partner shall make arrangements to make appropriate reasonable adjustments for candidates who require this, as agreed with the Exam Committee in advance	E
13	The Partner shall provide the Exam Committee with the results of the examination within 2 working days in an agreed format. Please describe this format, bearing in mind your response to requirements laid down in the next section of this bid	E
14	The Partner will have a Business Continuity Plan for all reasonably-foreseeable eventualities	E
15	All personal identifiable information is treated in accordance with GDPR requirements as applicable. The Partner takes responsibility to determine such requirements	E

Appendix 4: Application form

European Examination in Medical Microbiology

1.0 APPLICANT INFORMATION

Full Name: _____

Last *First*

Address: _____

Street Address *City*

_____

State *Country* *ZIP Code*

Phone: _____ *Email* _____

Position _____

Are currently a trainee in Medical Microbiology? YES ☐ NO ☐

If yes, how many months of training in Medical Microbiology will you have completed on the first day of May 2022? _____

Site(s) of Medical Microbiology training?

_____

_____

Are currently a Consultant in Medical Microbiology? YES ☐ NO ☐

If yes, how long will you have been working as a Consultant in Medical Microbiology on the first day of May 2022?

This exam is being delivered in collaboration with the European Board of Radiology

Site(s) of Medical Microbiology training?

_____

_____

2.0 EDUCATION

1. MEDICAL DEGREE

Address of University or College: _____

_____

_____

DATES:

From: _____ *To:* _____ *Did you graduate?* YES ☐ NO ☐

2. OTHER DEGREE/QUALIFICATION

Address of University or College: _____

_____

_____

DATES:

From: _____ *To:* _____ *Did you graduate?* YES ☐ NO ☐

3. OTHER DEGREE/QUALIFICATION

Address of University or College: _____

_____

_____

DATES:

From: _____ *To:* _____ *Did you graduate?* YES ☐ NO ☐

4. **OTHER DEGREE/QUALIFICATION**

Address of University or College: _____

_____

_____

DATES:

From: _____ *To:* _____ *Did you graduate?* YES ☐ NO ☐

3.0 REFERENCES:

Please list two referees who have supervised you during your training in Medical Microbiology.

Your Referees will need to confirm that you meet the eligibility criteria to sit this exam, as outlined in the document Terms and Conditions for Candidates applying to sit the European Examination in Medical Microbiology.

Full Name: _____ *Relationship:* _____

Position: _____ *Phone:* _____

Address: _____

Email: _____

I hereby certify that this applicant _____

Meets the eligibility criteria to sit the European Examination in Medical Microbiology,

Signed _____

Full Name: _____ *Relationship:* _____

Position: _____ *Phone:* _____

Address: _____

Email: _____

I hereby certify that this applicant _____

Meets the eligibility criteria to sit the European Examination in Medical Microbiology,

Signed _____

4.0 OTHER PROFESSIONAL EXAMINATION IN MEDICAL MICROBIOLOGY:

Have you sat any other professional examination in Medical Microbiology in your own country or another country e.g. FRCPath (UK).

YES ☐ NO ☐

If Yes

Could you please state what examination you sat?

_____

What date you took that examination?

_____

If no,

Do you plan to sit another professional examination in Medical Microbiology in the future e.g. FRCPath (UK) ?

YES ☐ NO ☐

If yes,

When do you plan to sit the exam? _____

5.0 AGREEMENT OF CONFIDENTIALITY

I agree with the following: any breach of confidentiality concerning the content of the examination will be reported to your National Society for Medical Microbiology/Professional Training Body/Organisation, which is represented in the National Notifying Authority of your country, as well as to your referees (trainer / staff member). In addition, such a breach may have consequences in accordance with regulations of the UEMS.

YES ☐ NO ☐

6.0 DISCLAIMER AND SIGNATURE

I have read the terms and conditions of participation in the European Examination in Medical Microbiology.

YES ☐ NO ☐

I accept the terms and conditions of participation in the European Examination in Medical Microbiology.

YES ☐ NO ☐

I certify that my answers are true and complete to the best of my knowledge.

Signature: _____ *Date:* _____

1. The candidate should register on the EBR website.
2. The application form should be downloaded, and completed in full
3. The application form must be submitted before the closing date, 20th April 2022, together with the following documents:

A scanned copy of the candidate's passport or I.D.

A scanned copy of this form, including the referees' signatures for certification of eligibility (see section 3.0, on page 3 and 4 of this form)

Candidates are responsible for the validity of all documents and the application data.

Supporting documents should be uploaded to this platform, according to the instructions on the registration website page.

Candidates are responsible for the validity of all documents and the application data.

Supporting documents should be uploaded to this platform, according to the instructions on the registration website page.

Appendix 5: Proposed Pathway to European Fellowship in MM, two options

Figure 1: Schema for EEMM and EFMM OPTION 1

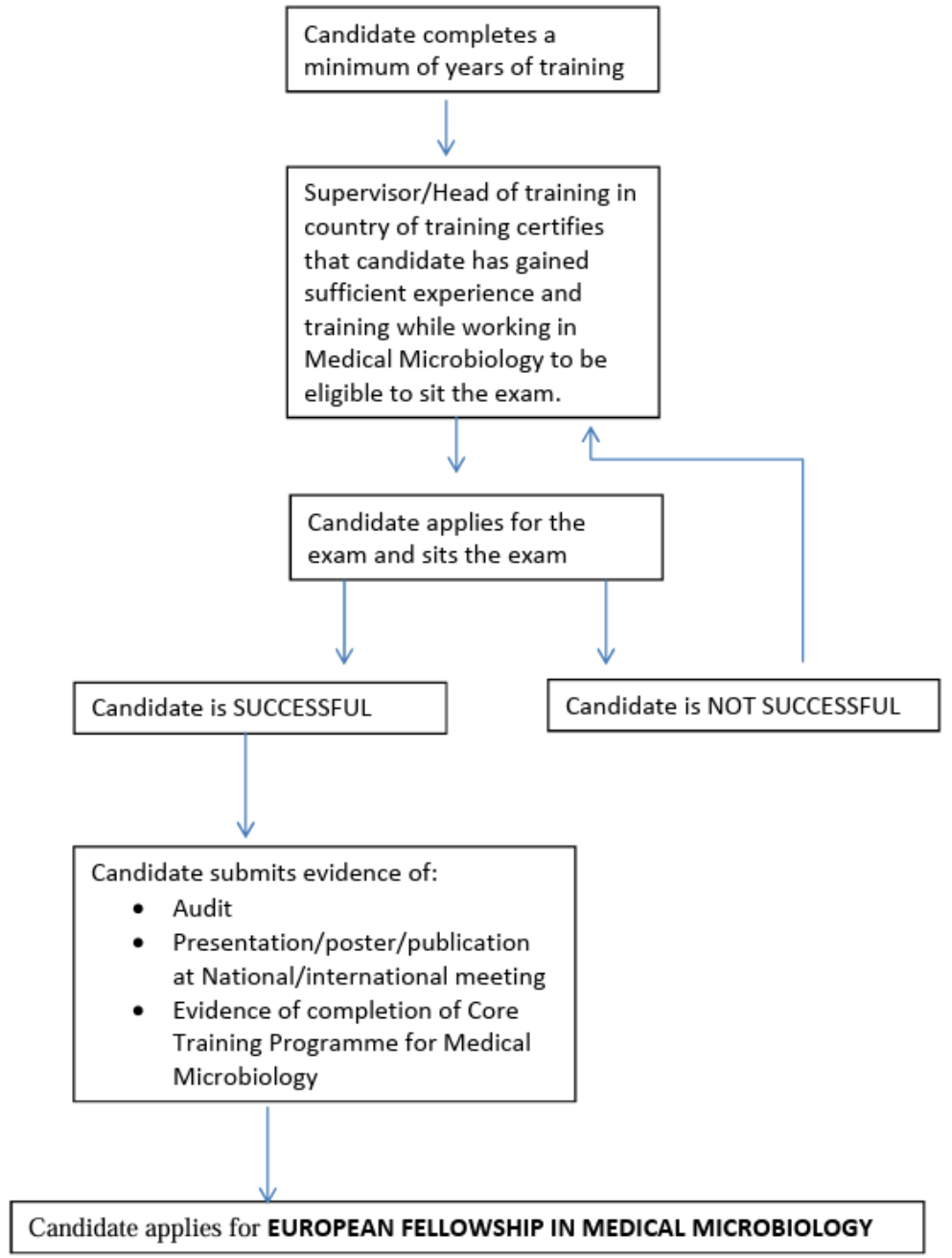
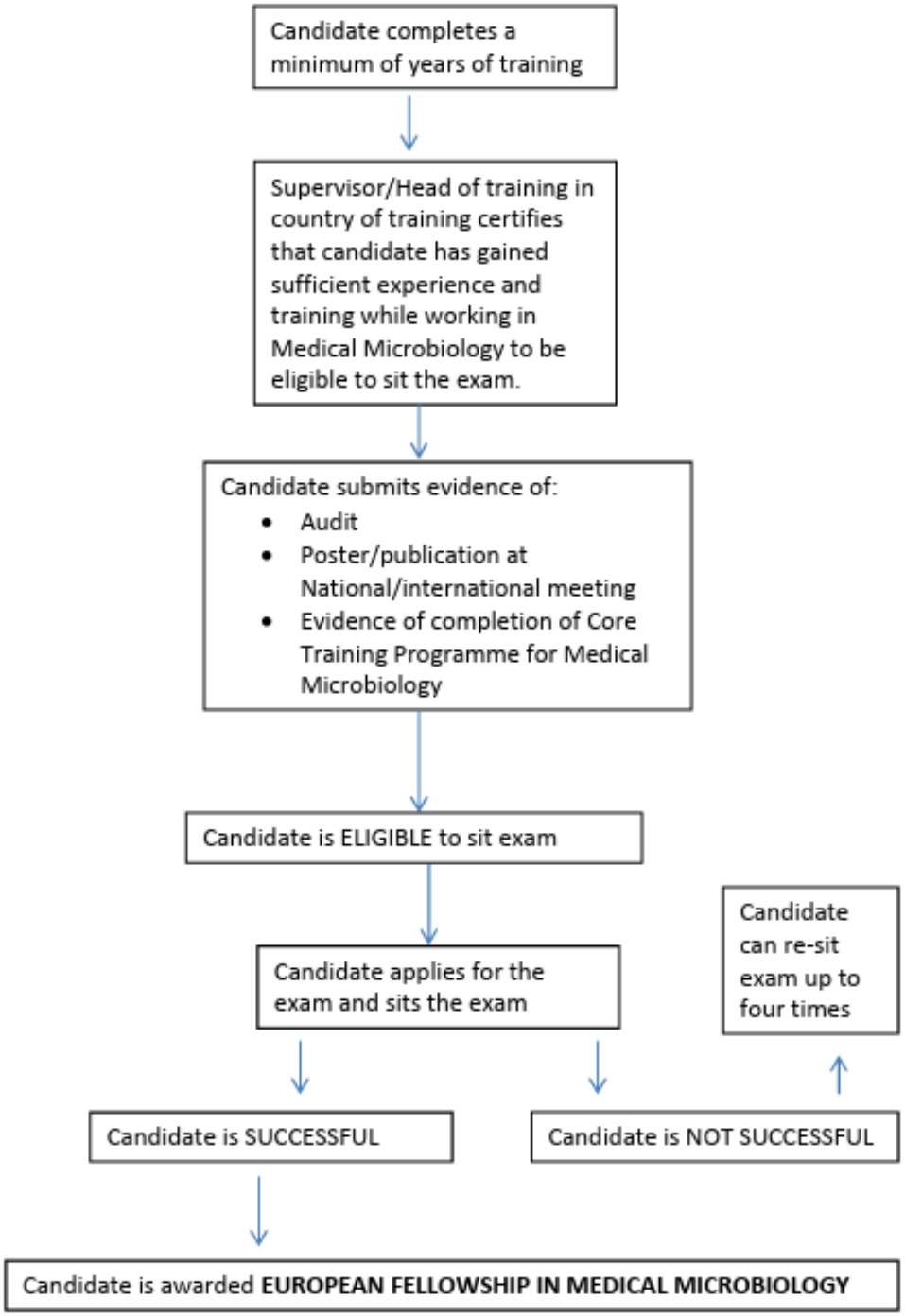


Figure 2: Schema for EEMM and EFMM OPTION 2



2018 Proposed Framework for European Microbiology Exam with additions for discussion. Version 2. Maeve

Appendix 6: Training Survey Questionnaire

* 1. I have received and read the Participant Information Leaflet

- ◯ Yes
- ◯ No

* 2. I agree to participate in this survey and consent to the data collected being used and stored as outlined in the Participant Information leaflet.

- ◯ Yes
- ◯ No

* 3. Which country are you from?

Part 1 UEMS-MM Questionnaire

Please provide information on the following aspects of medical microbiology training in your country

4. Do you have to be a doctor in order to train as a medical microbiology specialist in your country?

- ◯ Yes
- ◯ No

If no, please specify other eligible professions in your country

5. Do you have to complete a formal medical microbiology training programme in order to register as a medical microbiology specialist in your country?

- ◯ Yes
- ◯ No

If no, please briefly describe alternate pathways to becoming a registered medical microbiology specialist in your country.

6. In your country, is there a structured training programme for trainees who wish to become medical microbiology specialists?

- ◯ Yes
- ◯ No
- ◯ Other - please describe

7. If yes to question 5, what is the title of the medical microbiology training programme in your country?

8. What is the current status of medical microbiology training in your country? *Eg primary specialty, subspeciality etc*

9. Who is the overseeing training body/national authority responsible for the organisation and delivery of medical microbiology training?

10. Is there a designated national clinical lead for the medical microbiology training programme? If yes, who fulfills this role?

- ◯ Yes
- ◯ No

If yes, who fulfils this role?

11. Do trainees have to spend a specific amount of time working in clinical medicine before entering medical microbiology training?

- ◯ Yes
- ◯ No

If yes, please describe

12. What are the minimum entry requirements for training in medical microbiology in your country?

13. What is the duration of medical microbiology training?

- ◯ 1 year
- ◯ 2 years
- ◯ 3 years
- ◯ 4 years
- ◯ 5 years
- ◯ 6 years
- ◯ Other (please specify)

14. How many trainees are currently on the medical microbiology training programme or currently in training to become a medical microbiology specialist?

- ◯ <25
- ◯ 25-50
- ◯ 51-75
- ◯ 76-100
- ◯ >100
- ◯ Other (please specify)

15. What is the average annual intake of new trainees?

- ◯ <10
- ◯ 10-20
- ◯ 21-30
- ◯ 31-40
- ◯ 41-50
- ◯ >50
- ◯ Other (please specify)

Organisation of medical microbiology training

16. Is the medical microbiology training in your country national or regional?
(E.g. do the trainees rotate through different training sites nationally throughout the duration of training, or do they complete it all in one site/hub of local sites?)

- ◯ Trainees stay in one site for duration of training
- ◯ Trainees rotate sites within a regional hub of sites
- ◯ Trainees rotate sites nationally
- ◯ Other (please describe)

17. How long is spent in each training site?

- ◯ Less than 6 months
- ◯ 6 months
- ◯ 1 year
- ◯ 2 years
- ◯ Total duration of training
- ◯ Other (please describe)

18. Are there any mandatory training sites?

- ◯ Yes
- ◯ No

If yes, please describe

19. Does each trainee have an assigned trainer or supervisor with responsibility for their training?

- ◯ Yes
- ◯ No
- ◯ Other (please describe)

20. If the answer is yes to question 16 above, are trainees reassigned new trainers/supervisors for each year of training?

- ◯ Each trainee is assigned a new trainer with each year of training
- ◯ Each trainee has the same trainer throughout the training programme
- ◯ Other (please describe)

21. Is there a training curriculum with specific learning outcomes?

- ◯ Yes
- ◯ No

If yes, can you provide a link to the document?

22. Please indicate below what percentage of your trainees training time is dedicated to each of the following areas:

	No dedicated time (0%)	Infrequent dedicated time (1-5%)	6-25%	26-50%	51-75%	Frequent dedicated time (76-95%)	Almost all training time (96-99%)	All training time (100%)
Practical laboratory benchwork	○	○	○	○	○	○	○	○
Laboratory management experience	○	○	○	○	○	○	○	○
Clinical liaison and infection management advice	○	○	○	○	○	○	○	○
Infection prevention and control experience	○	○	○	○	○	○	○	○
Antimicrobial stewardship experience	○	○	○	○	○	○	○	○

Other - please describe any other areas that your trainees spend dedicated time on

23. Are out of clinical programme experience (OCPE) years available to the trainees?
(E.g. to pursue research/teaching/particular subspecialities of interest that may not otherwise be available to the trainee)

- ○ Yes - for research opportunities
- ○ Yes - for teaching/lecturing opportunities
- ○ Yes - for subspeciality experience
- ○ Yes - for all of the above
- ○ Yes - other (please provide details below)
- ○ No - an OCPE year is not available
- ○ Other: please provide details

24. If an out of clinical programme year is available to trainees, is this time in addition to, or counted towards, the usual duration of training?

- ◯ The out of clinical programme year is in addition to the usual training years
- ◯ The out of clinical programme year is counted as one of the usual training years
- ◯ No out of clinical programme year available
- ◯ Other (please describe)

25. Is the option of part-time or flexible training available to your trainees?

- ◯ Yes
- ◯ No

If yes, please provide details
e.g. eligibility criteria, duration of part-time/flexible training available, provision of funding for part-time training etc

Documentation of required training activities

26. Is there a mandatory logbook or record where each trainee must document required training activities?

- ◯ Yes - paper record
- ◯ Yes - electronic record
- ◯ Yes - other, please describe below
- ◯ There is no mandatory record of training activities
- ◯ Other (please describe)

Which of the following are training activities that must be recorded by the trainees?

Please select all training activities that must be recorded from the questions below

27. Are trainees required to create a training plan/personal goals plan?

- ◯ Required once per year
- ◯ Required once per training programme
- ◯ Not required
- ◯ Other (please describe)

28. Are trainees required to keep a record of attendance at any of the following department/hospital educational sessions?
Please select all that apply

- ☐ Journal clubs
- ☐ Grand rounds
- ☐ Record not required
- ☐ Other (please describe)

29. Are trainees required to deliver and record any of the following teaching sessions?
Please select all that apply

- ☐ Teaching sessions delivered to undergraduate students e.g. lectures/tutorials/clinical teaching
- ☐ Teaching sessions delivered to postgraduate students e.g. lectures/tutorials/clinical teaching
- ☐ Not required
- ☐ Other (please describe)

30. Are trainees required to record audits they have completed throughout the training programme?

- ◯ Required to record a minimum of one per year
- ◯ Required to record a minimum of one per training programme
- ◯ Not required
- ◯ Other (please describe)

31. Are trainees required to record any of the following research activities?
Please select all that apply

- ☐ Record of at least one publication
- ☐ Record of at least one presentation at a national/international meeting
- ☐ Record of attendance at national/international meetings
- ☐ Not required
- ☐ Other (please describe)

32. Are trainees required to keep a record of their attendance at any of the following hospital committee meetings?
Please select all that apply

- ☐ Infection control committee meetings
- ☐ Lab management meetings
- ☐ Not required
- ☐ Other (please describe)

33. Are trainees required to record examples of clinical liaison and case discussions with any of the following specialities?
Please select all that apply

- [ ] General practitioners
- [ ] Haematology
- [ ] Oncology
- [ ] Acute medicine
- [ ] Surgery
- [ ] Orthopaedics
- [ ] Paediatrics
- [ ] Obstetrics
- [ ] Public Health
- [ ] Food and water microbiology
- [ ] Not required
- [ ] Other (please specify)

34. Are trainees required to record examples of experience gained in any of the following infection control activities?
Please select all that apply

- [ ] Surveillance
- [ ] Outbreak management
- [ ] Disinfection
- [ ] Sterilisation
- [ ] Hospital building issues
- [ ] Environmental issues
- [ ] Inoculation injuries
- [ ] Not required
- [ ] Other (please specify)

35. Are trainees required to record demonstration of the following practical laboratory procedures and skills?
Please select all that apply

- [ ] Culture (bacteria, fungi, mycobacteria) and reading of plates
- [ ] Antimicrobial susceptibility testing
- [ ] Identification
- [ ] Microscopy
- [ ] Nucleic acid detection
- [ ] Sample preparation
- [ ] Serological testing
- [ ] Not required
- [ ] Other (please specify)

36. Are trainees required to keep a record of relatively unusual clinical cases that they were involved in throughout their training?

- ( ) Yes
- ( ) No
- ( ) Other (please specify)

37. Are trainees required to keep a record of ward rounds carried out in any of the following areas?
Please select all that reply

- [ ] Intensive care unit rounds
- [ ] Antimicrobial stewardship rounds
- [ ] Not required
- [ ] Other (please describe)

38. Are trainees required to record their involvement in the development of practice guidelines or policies during the training programme?
Please select all that apply

- ◯ Yes
- ◯ No
- ◯ Other (please describe)

39. Are trainees required to record any of the following examples of management experience?
Please select all that apply

- ☐ Workload management
- ☐ Laboratory accreditation
- ☐ Complaints procedures
- ☐ Recruitment
- ☐ Not required
- ☐ Other (please describe)

40. Are there any other components of your logbook or record of training activities not covered in the above?
Please describe

Assessment methods

41. Are trainees required to meet with their assigned trainers/supervisors at regular intervals throughout training in order to evaluate their training progress?

- ◯ Quarterly meetings are required
- ◯ Annual meetings are required
- ◯ No meetings are required
- ◯ Other (please describe)

42. Are trainees required to complete any of the following formative assessments under the supervision of their trainers during their training?
Please select all that apply

- ☐ Mini clinical experiences (A mini-CEX is based on a 15 minute observation of a single interaction and is useful for assessing clinical skills, attitudes and behaviours that are essential to high quality patient care)
- ☐ Directly observed procedural skills (DOPS - the trainee performs a procedure e.g. laboratory skills, and is observed by an experienced and knowledgeable assessor who reviews the trainee's performance and provides feedback)
- ☐ Case-based discussions (CBDs - used to explore the Trainee's clinical assessments, investigations, reasoning, judgement, decision-making and management skills.)
- ☐ None required
- ☐ Other (please describe)

43. Are trainees required to pass a specific national or international exit examination in order to register as a specialist?

- ◯ Yes
- ◯ No

if yes, please list the required examinations

44. Aside from the above, are there any other forms of assessment used to measure trainees' progress throughout the scheme?
Please provide details.

Mandatory courses and study days

45. Are there mandatory training courses that the trainees must attend?

- ◯ Yes
- ◯ No

If yes, please list/describe

46. Are there mandatory study days that the trainees must attend?

- ◯ Yes
- ◯ No

If yes how many per year?
Who is responsible for the organisation and delivery of the study days?

47. Are trainees required to complete any other additional qualifications by end of training?

- ◯ Yes
- ◯ No

If yes, please describe

On-call commitments

48. Do the trainees have mandatory 'on-call' or 'out-of-hours' commitments?

- ◯ Yes
- ◯ No
- ◯ Other (please describe)

49. If yes, what is the nature of the 'on-call' or 'out-of-hours' commitment?

Part 2 UEMS-MM Questionnaire - optional

Identification of national authorities responsible for aspects of medical microbiology specialist training in Europe.

In your country, please provide details of the responsible national authority for each aspect of medical microbiology specialist training listed below, if available to you.

50. Please provide details of the responsible national authority for **training curriculum content:**

Name in local language	
Name in English	
Address	
Address 2	
City/Town	
State/Province	
ZIP/Postal Code	
Country	
Email Address	
Phone Number (incl country code)	

51. Please provide details of the responsible national authority for **training curriculum delivery (oversight of standards of delivery of training):**

Name in local language	
Name in English	
Address	
Address 2	
City/Town	
State/Province	
ZIP/Postal Code	
Country	
Email Address	
Phone Number (incl national code)	

52. Please provide details of the responsible national authority for **allocation of training resources and positions (including funding):**

Name in local language	
Name in English	
Address	
Address 2	
City/Town	
State/Province	
ZIP/Postal Code	
Country	
Email Address	
Phone Number (incl national code)	

53. Please provide details of the responsible national authority for **licensing of medical microbiology specialists:**

Name in local language	
Name in English	
Address	
Address 2	
City/Town	
State/Province	
ZIP/Postal Code	
Country	
Email Address	
Phone Number (incl national code)	

54. Please provide details of the **national specialist society delegating UEMS-MM representation:**

Name in English	
Name in local language	
Address	
Address 2	
City/Town	
State/Province	
ZIP/Postal Code	
Country	
Email Address	
Phone Number (incl national code)	

References

Adam, S. (2004). A consideration of the nature, role, application and implications for European education of employing 'learning outcomes' at the local, national and international levels.

Alexander, E., N. Osman, J. Walling and V. Mitchell (2012). "Variation and Imprecision of Clerkship Grading in U.S. Medical Schools." Academic medicine : journal of the Association of American Medical Colleges **87**: 1070-1076.

Allerberger, F. (2021). "Training in clinical microbiology and infectious diseases in Europe." Clin Microbiol Infect **27**(11): 1575.

Alsulimani, L. K. (2021). "The feasibility of simulation-based high-stakes assessment in emergency medicine settings: A scoping review." J Educ Health Promot **10**: 441.

Anderson, L., D. Krathwohl, P. Airasian and K. Cruikshank (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.

Andreou, V., S. Peters, J. Eggermont, J. Wens and B. Schoenmakers (2021). "Remote versus on-site proctored exam: comparing student results in a cross-sectional study." BMC Med Educ **21**(1): 624.

Angoff, W. H. (1971). Scale, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement. Washington, DC, American Council on Education.

Arain, M., M. J. Campbell, C. L. Cooper and G. A. Lancaster (2010). "What is a pilot or feasibility study? A review of current practice and editorial policy." BMC Medical Research Methodology **10**(1): 67.

Arnò, S., A. Galassi, M. Tommasi, A. Saggino and P. Vittorini (2021). "State-of-the-Art of Commercial Proctoring Systems and Their Use in Academic Online Exams." International Journal of Distance Education Technologies **19**: 41-60.

Asikainen, H., A. Parpala, S. Lindblom-Ylänne, G. Vanthournout and L. Coertjens (2014). "The Development of Approaches to Learning and Perceptions of the Teaching-Learning Environment During Bachelor Level Studies and Their Relation to Study Success." Higher Education Studies **4**: 24-36.

Avenia-Tapper, B. and L. Llosa (2015). "Construct Relevant or Irrelevant? The Role of Linguistic Complexity in the Assessment of English Language Learners' Science Knowledge." Educational Assessment **20**: 95-111.

Bandaranayake, R. C. (2008). "Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37." Med Teach **30**(9-10): 836-845.

Bandaranayake, R. C. (2008). "Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37." Medical Teacher **30**(9-10): 836-845.

Barnett, R., G. Parry and K. Coate (2001). "Conceptualising Curriculum Change." Teaching in Higher Education - TEACH HIGH EDUC **6**: 435-449.

Barr, H., I. Koppel, S. Reeves, M. Hammick and D. Freeth (2006). "Effective Interprofessional Education: Argument, Assumption and Evidence." Effective Interprofessional Education: Argument, Assumption and Evidence.

Battey, H., B. Doran, A. Flood, J. Nussbaum, T. Seto, S. Srisatidnarakul, B. Tegtmeier and S. Dadwal (2023). "The COVID-19 infection control response at a large stand-alone comprehensive cancer center in Los Angeles County." Cancer Rep (Hoboken) **6**(1): e1669.

Baumann, M., J. W. H. Leer, O. Dahl, W. De Neve, R. Hunter, R. Rampling and C. Verfaillie (2004). "Updated European core curriculum for radiotherapists (radiation oncologists). Recommended curriculum for the specialist training of medical practitioners in radiotherapy (radiation oncology) within Europe." Radiotherapy and Oncology **70**(2): 107-113.

Bejar, I. (2008). Standard Setting: What Is It? Why Is It Important?

Bell, R. (1973). Thinking about the curriculum. Milton Keyes, Open University Press.

Belovarac, B. J., S. R. Zabar, D. Warfield, M. A. Bannan and A. V. Rapkiewicz (2021). "The OSPE." Am J Clin Pathol **155**(3): 324-332.

Belson, W. (1981). The Design and Understanding of Survey Questions. Aldershot, UK, Gower.

Beretvas, S. (2004). "Comparison of Bookmark Difficulty Locations Under Different Item Response Models." Applied Psychological Measurement - APPL PSYCHOL MEAS **28**: 25-47.

Bhakta, B., A. Tennant, M. Horton, G. Lawton and D. Andrich (2005). "Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education." BMC Medical Education **5**(1): 9.

Biggs, J. (1999). "Teaching for quality learning at university (2, ed.) Buckingham."

Biggs, J. (2003). "Aligning teaching and assessing to course objectives." International Conference on Teaching and Learning in Higher Education: New trend and innovations **2**.

Biggs, J. (2003). Teaching for Quality Learning at University.

Biggs, J. (2014). "Constructive alignment in University Teaching." HERDSA Rev High Educ **1**: 5-22.

Bloom, B. S. and D. R. Krathwohl (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals, Longmans, Green.

Boet, S., S. Sharma, J. Goldman and S. Reeves (2012). "Review article: medical education research: an overview of methods." Can J Anaesth **59**(2): 159-170.

Boone, W. J. (2016). "Rasch Analysis for Instrument Development: Why, When, and How?" CBE Life Sci Educ **15**(4).

Borsboom, D., G. Mellenbergh and J. Heerden (2004). "The Concept of Validity." Psychological review **111**: 1061-1071.

Boursicot, K. (2006). "Setting Standards in a Professional Higher Education Course: Defining the Concept of the Minimally Competent Student in Performance-Based Assessment at the Level of Graduation from Medical School." Higher Education Quarterly **60**(1): 74-90.

Bowe, S. N., R. A. Bly, X. Wang and M. E. Whipple (2022). "Racial and Ethnic Differences in Resident Selection in 11 Specialties, 2013-2018." JAMA **327**(24): 2450-2452.

Brauer, D. G. and K. J. Ferguson (2015). "The integrated curriculum in medical education: AMEE Guide No. 96." Med Teach **37**(4): 312-322.

Brearley, S. (1995). "Harmonisation of specialist training in Europe: is it a mirage?" Bmj **311**(7000): 297-299.

Brennan, R. (1984). Estimating the dependability of the scores. In A guide to criterion-referenced test construction (Ed.) Ronald A Berk. Baltimore, The Johns Hopkins University Press.

Brown, G., J. Bull and M. Pendlebury (1997). Assessing Student Learning in Higher Education. London, Routledge.

Buckendahl, C. W., R. W. Smith, J. C. Impara and B. S. Plake (2002). "A Comparison of Angoff and Bookmark Standard Setting Methods." Journal of Educational Measurement **39**(3): 253-263.

Bugge, C., B. Williams, S. Hagen, J. Logan, C. Glazener, S. Pringle and L. Sinclair (2013). "A process for Decision-making after Pilot and feasibility Trials (ADePT): Development following a feasibility study of a complex intervention for pelvic organ prolapse." Trials **14**: 353.

Burgard, T., M. Bošnjak and N. Wedderhoff (2020). "Response rates in online surveys with affective disorder participants: A meta-analysis of study design and time effects between 2008 and 2019." Zeitschrift für Psychologie **228**(1): 14-24.

Burmeister, L. F. (2003). "Principles of successful sample surveys." Anesthesiology **99**(6): 1251-1252.

Caruana, E. J., A. Patel, S. Kendall and S. Rathinam (2020). "Impact of coronavirus 2019 (COVID-19) on training and well-being in subspecialty surgery: A national survey of cardiothoracic trainees in the United Kingdom." J Thorac Cardiovasc Surg **160**(4): 980-987.

Case, S. and D. Swanson (2002). "Constructing Written Test Questions For the Basic and Clinical Sciences." National Board of Examiners.

Case, S. M. and D. B. Swanson (1993). "Extended-matching items: A practical alternative to free-response questions." Teaching and Learning in Medicine **5**(2): 107-115.

çetin, S. and S. Gelbal (2013). "A Comparison of Bookmark and Angoff Standard Setting Methods." Educational Sciences: Theory and Practice **13**: 2169-2175.

Chai, D. (2016). Automated marking of printed multiple choice answer sheets.

Chandra, S., A. Kusum, D. S. Gaur and H. Chandra (2022). "Analytical and Post Analytical Phase of an ISO 15189:2012 Certified Cytopathology Laboratory-A Five Year Institutional Experience." J Cytol **39**(1): 37-43.

Check, J. and R. Schutt (2012). Research Methods in Education.

Chirumamilla, A., G. Sindre and A. Nguyen Duc (2020). "Cheating in e-exams and paper exams: the perceptions of engineering students and teachers in Norway." Assessment & Evaluation in Higher Education.

Choi, B., L. Jegatheeswaran, A. Minocha, M. Alhilani, M. Nakhoul and E. Mutengesa (2020). "The impact of the COVID-19 pandemic on final year medical students in the United Kingdom: a national survey." BMC Med Educ **20**(1): 206.

Cilliers, F. J., L. W. Schuwirth, H. J. Adendorff, N. Herman and C. P. van der Vleuten (2010). "The mechanism of impact of summative assessment on medical students' learning." Adv Health Sci Educ Theory Pract **15**(5): 695-715.

Cizek, G. and M. Bunch (2007). "Standard setting: A guide to establishing and evaluating performance standards on tests."

Cizek, G. J. (1996). "Standard-Setting Guidelines." Educational Measurement: Issues and Practice **15**(1): 13-21.

Cizek, G. L. (2001). "More unintended consequences of high-stakes testing." Educational Measurement: Issues and Practice **20**(4): 19-27.

Clauser, B. E., P. Harik, M. J. Margolis, I. C. McManus, J. Mollon, L. Chis and S. Williams (2009). "An empirical examination of the impact of group discussion and examinee performance information on judgments made in the angoff standard-setting procedure." Applied Measurement in Education **22**(1): 1-21.

Cochran-Smith, M. and S. L. Lytle (2001). Beyond certainty: taking an inquiry stance on practice. In: Lieberman, A.; Miller, L. (Eds). Teachers caught in the action: professional development that matters. New York, Teachers College Press.

Coderre, S., W. Woloschuk and K. McLaughlin (2009). "Twelve tips for blueprinting." Med Teach **31**(4): 322-324.

Cohen-Schotanus, J. and C. P. van der Vleuten (2010). "A standard setting method with the best performing students as point of reference: practical and affordable." Med Teach **32**(2): 154-160.

Cohen-Schotanus, J. and C. P. M. Van der Vleuten (1996). "Een betere cesuur bij tentamens [A better standard setting method for written tests]." Onderzoek van Onderwijs **September**: 54-55.

Cole, J. and A. Kamondi (2023). "A proposal for harmonizing clinical neurophysiology training in the Europe, Middle East and Africa Chapter of the International Federation of Clinical Neurophysiology." Clinical Neurophysiology **156**: 76-85.

Cole, J., A. Kamondi, P. T. Kimaid and N. Shahrizaila (2022). "Training and education practice in the Europe, Middle East and Africa, Latin America and Asia Oceania chapters, IFCN; an international survey." Clinical Neurophysiology Practice **7**: 120-126.

Commission, E., Y. Directorate-General for Education, Sport and Culture (2015). ECTS users' guide 2015, Publications Office of the European Union.

Commission, E. P. (2012). "ESCMID Definition of Geographic areas." Retrieved March 2024, 2024.

Communiqué, B. (2003). "“Realising the European Higher Education Area”." Communiqué of the Conference of Ministers responsible for Higher Education.

Communiqué, B. (2005). "The European Higher Education Area -Achieving the Goals." Bologna Procss.

Cook, C., F. Heath and R. L. Thompson (2000). "A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys." Educational and Psychological Measurement **60**(6): 821-836.

Cook, T. D. and D. T. Campbell (1979). Quasi-experimentation: Design & Analysis Issues for Field Settings, Houghton Mifflin.

Coombes, L., M. Roberts, D. Zahra and S. Burr (2016). "Twelve tips for assessment psychometrics." Med Teach **38**(3): 250-254.

Cotton, D., P. Cotton and J. Shipway (2023). "Chatting and cheating. Ensuring academic integrity in the era of ChatGPT. EdArXiv." Preprint posted online January **10**.

Cox, M., D. M. Irby and R. M. Epstein (2007). "Assessment in medical education." New England Journal of Medicine **356**(4): 387-396.

Cronbach, L. (1971). Test validation. In R. L. Thorndike (Ed.) Educational measurement. Washington DC, American Council on Education.

Cronbach, L. and P. Meehl (1955). "Construct Validity in Psychological Test." Psychol Bull **52**: 281-302.

D.M., L., H. C. Mitzel and D. R. Green (1996). Standard Setting: A Bookmark Approach. In D. R. Green (Chair), IRT-Based Standard-Setting Procedures Utilizing Behavioral Anchoring. 1996 Council of Chief State School Officers 1996 National Conference on Large Scale Assessment. Phoenix, AZ.

Dave, R. H. (1970). Psychomotor levels. In R. J. Armstrong (Ed.), Developing and writing behavioral objectives, Tucson, AZ: Educational Innovators Press.

Dawber, T. (2002). Running head : COGNITIVE EXPERIENCE OF BOOKMARK PARTICIPANTS The Cognitive Experience of Bookmark Standard Setting Participants.

Dayal, A., D. M. O'Connor, U. Qadri and V. M. Arora (2017). "Comparison of Male vs Female Resident Milestone Evaluations by Faculty During Emergency Medicine Residency Training." JAMA Intern Med **177**(5): 651-657.

De Champlain, A. (2010). "A primer on classical test theory and item response theory for assessment in medical education." Medical education **44**: 109-117.

De Champlain, A. F. (2018). "Standard setting methods in medical education: high-stakes assessment." Understanding medical education: Evidence, theory, and practice: 347-359.

De Vet, H., C. Terwee, L. Mokkink and D. Knol (2011). "Measurement in Medicine: A Practical Guide." Measurement in Medicine: A Practical Guide: 1-338.

Desselle, S. (2005). "Construction, Implementation, and Analysis of Summated Rating Attitude Scales." American Journal of Pharmaceutical Education **69**.

Dickstein, Y., R. Nir-Paz, C. Pulcini, B. Cookson, B. Beović, E. Tacconelli, D. Nathwani, R. Vatcheva-Dobrevska, J. Rodríguez-Baño, M. Hell, H. Saenz, L. Leibovici and M. Paul (2016). "Staffing for infectious diseases, clinical microbiology and infection control in hospitals in 2015: results of an ESCMID member survey." Clinical Microbiology and Infection **22**(9): 812.e819-812.e817.

Dijkstra, J., C. P. Van der Vleuten and L. W. Schuwirth (2010). "A new framework for designing programmes of assessment." Adv Health Sci Educ Theory Pract **15**(3): 379-393.

Dillman, D. (2000). Mail and Internet Surveys: The Tailored Design Method.

Dinsmore, D. L. and P. A. Alexander (2012). "A Critical Discussion of Deep and Surface Processing: What It Means, How It Is Measured, the Role of Context, and Model Specification." Educational Psychology Review **24**(4): 499-567.

Donald, G. (2018). "A Brief Summary of Pilot and Feasibility Studies: Exploring Terminology, Aims, and Methods." European Journal of Integrative Medicine **24**.

Downing, S. (2004). "Reliability: On the reproducibility of assessment data." Medical education **38**: 1006-1012.

Downing, S., A. Tekian and R. Yudkowsky (2006). "RESEARCH METHODOLOGY: Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education." Teaching and learning in medicine **18**: 50-57.

Downing, S. M. (2003). "Validity: on meaningful interpretation of assessment data." Med Educ **37**(9): 830-837.

Downing, S. M. (2004). "Reliability: on the reproducibility of assessment data." Med Educ **38**(9): 1006-1012.

Downing, S. M., A. Tekian and R. Yudkowsky (2006). "Procedures for establishing defensible absolute passing scores on performance examinations in health professions education." Teach Learn Med **18**(1): 50-57.

Downing, S. M. and R. Yudkowsky (2009). Assessment in Health Professions Education, Routledge.

Doyle, M., B. Boyle, C. Brennan, J. Holland, A. Mifsud, M. Hell, F. van Tiel and T. M. Leegaard (2021). "Specialist training in medical microbiology across Europe in 2021-an update on the actual training situation based on a survey." Clin Microbiol Infect **27**(11): 1576-1580.

Draugalis, J. R., S. J. Coons and C. M. Plaza (2008). "Best practices for survey research reports: a synopsis for authors and reviewers." Am J Pharm Educ **72**(1): 11.

Dunn, W. R., D. D. Hamilton and R. M. Harden (1985). "Techniques of identifying competencies needed of doctors." Med Teach **7**(1): 15-25.

Durkan, M. (2023). CESMA Picking your exam. CESMA. Brussels, Belgium.

Dzara, K. and H. Gooding (2022). "A Guide to Educational Pyramids Commonly Used in Medical Education Programs." Acad Med **97**(2): 313.

Ebel, R. L. (1972). Essentials of Educational Measurement, Prentice-Hall.

Ebel, R. L. (1979). Essentials of Educational Measurement, Prentice-Hall.

Ebel, R. L. (1983). "The Practical Validation of Tests of Ability." Educational Measurement: Issues and Practice **2**(2): 7-10.

ECTS (2005) "ECTS Users' Guide."

Education, A. C. f. G. M. (2010). Key considerations for selecting assessment instruments and implementing assessment systems.

Egan, K. (2001). Validity and defensibility of cutscores established by the Bookmark standard setting method. 2001 Council of Chief State School Officers Conference on Large-Scale Assessment. Houston.

Egan, K., S. Ferrara, C. Schneider and K. Barton (2009). "Writing Performance Level Descriptors and Setting Performance Standards for Assessments of Modified Achievement Standards: The Role of Innovation and Importance of Following Conventional Practice." Peabody Journal of Education **84**: 552-577.

EHEA. (1998). "SORBONNE DECLARATION." Bologna Process.

EHEA. (1999). "The Bologna Declaration." The Bologna Process Retrieved March 2024, 2024.

Elfaki, O. and K. Salih (2015). "Comparison of Two Standard Setting Methods in a Medical Students MCQs Exam in Internal Medicine." American Journal of Medicine and Medical Sciences **5**(4): 164-167.

Elfaki, O. and K. e. Salih (2018). "Comparison of Two Standard Setting Methods in a Medical Students MCQs Exam in Internal Medicine."

Ellis, R., D. Scrimgeour and P. Brennan (2021). "The personal cost of postgraduate medical exams: Are we asking too much of trainees?" BMJ: British Medical Journal **https://blogs.bmj.com/bmj/2021/02/02/the-personal-cost-of-postgraduate-medical-exams-are-we-asking-too-much-of-trainees/**.

Elsalem, L., N. Al-Azzam, A. A. Jum'ah and N. Obeidat (2021). "Remote E-exams during Covid-19 pandemic: A cross-sectional study of students' preferences and academic dishonesty in faculties of medical sciences." Ann Med Surg (Lond) **62**: 326-333.

ESG. (2015). "The Standards and Guidelines for Quality Assurance in the European Higher Education Area " Bologna Process.

EU, C. o. t. (1993). Council Directive 93/16/EEC of 5 April 1993 to facilitate the free movement of doctors and the mutual recognition of their diplomas, certificates and other evidence of formal qualifications. Council Directive 93/16/EEC. P. o. o. t. E. Union, EU.

EUR-Lex. (2005). "Directive 2005/36/EC - Recognition of Professional qualifications." EC Directives Retrieved March 2024, 2024.

Fawns, T. and S. Schaepkens (2022). "A Matter of Trust: Online Proctored Exams and the Integration of Technologies of Assessment in Medical Education." Teach Learn Med **34**(4): 444-453.

Fehrmann, M. L., D. J. Woehr and W. Arthur (1991). "The Angoff cutoff score method: The impact of frame-of-reference rater training." Educational and Psychological Measurement **51**(4): 857-872.

Feletti, G. I. (1980). "Reliability and validity studies on modified essay questions." J Med Educ **55**(11): 933-941.

Felt, J. M., R. Castaneda, J. Tiemensma and S. Depaoli (2017). "Using Person Fit Statistics to Detect Outliers in Survey Research." Front Psychol **8**: 863.

Fischer, L., T. Rohm, C. Carstensen and T. Gnambs (2021). "Linking of Rasch-Scaled Tests: Consequences of Limited Item Pools and Model Misfit." Frontiers in Psychology **12**.

French, S., A. Dickerson and R. A. Mulder (2024). "A review of the benefits and drawbacks of high-stakes final examinations in higher education." Higher Education **88**(3): 893-918.

Freunberger, R., C. Luger-Bazinger and U. Itzlinger-Bruneforth (2013). Linking the Austrian Standards-Based Test for English to the CEFR: The Standard-Setting Process.

Friedman Ben-David, M. (2000). "AMEE Guide No. 18: Standard setting in student assessment." Medical Teacher - MED TEACH **22**: 120-130.

Fry, H., S. Ketteridge and S. Marshall (2009). "A Handbook for Teaching and Learning in Higher Education: Enhancing Academic Practice." http://lst-iiep.iiep-unesco.org/cgi-bin/wwwi32.exe/[in=epidoc1.in]/?t2000=024772/(100).

García-Pérez, M. A., C. Amaya and Á. Otero (2007). "Physicians' migration in Europe: an overview of the current situation." BMC Health Services Research **7**(1): 201.

Gatawa, B. S. M. (1990). The Politics of the School Curriculum: An Introduction. Harare, Jongwe Press.

Gronlund, N. E. and R. L. Linn (1990). Measurement and Evaluation in Teaching, Macmillan.

Guagnano, M. T., D. Merlitti, M. R. Manigrasso, V. Pace-Palitti and S. Sensi (2002). "New medical licensing examination using computer-based case simulations and standardized patients." Acad Med **77**(1): 87-90.

Guangul, F. M., A. H. Suhail, M. I. Khalit and B. A. Khidhir (2020). "Challenges of remote assessment in higher education in the context of COVID-19: a case study of Middle East College." Educational Assessment, Evaluation and Accountability **32**(4): 519-535.

Hambleton, R. and M. Pitoniak (2006). "Setting Performance Standards." Educational Measurement: 433-470.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ, US, Lawrence Erlbaum Associates Publishers**:** 89-116.

Hamdy, H. (2006). "Blueprinting for the assessment of health care professionals." The Clinical Teacher **3**(2): 175-179.

Hansche, L. N. (1998). Handbook for the Development of Performance Standards: Meeting the Requirements of Title I.

Harden, R. M. (1986). "Approaches to curriculum planning." Medical Education **20**(5): 458-466.

Harden, R. M. (1986). "Ten questions to ask when planning a course or curriculum." Medical Education **20**.

Harden, R. M., J. R. Crosby and M. H. Davis (1999). "AMEE Guide No. 14: Outcome-based education: Part 1-An introduction to outcome-based education." Medical Teacher **21**: 7-14.

Harden, R. M., M. H. Davis and J. R. Crosby (1997). "The new Dundee medical curriculum: a whole that is greater than the sum of the parts." Med Educ **31**(4): 264-271.

Harden, R. M. and F. A. Gleeson (1979). "Assessment of clinical competence using an objective structured clinical examination (OSCE)." Med Educ **13**(1): 41-54.

Hargreaves, A. (1989). Curriculum and Assessment Reform. Milton Keynes,

UK, Open University Press.

Harrison, D. L. and J. R. Draugalis (1997). "Evaluating the results of mail survey research." J Am Pharm Assoc (Wash) **Ns37**(6): 662-666.

Harrow, A. J. (1972). A Taxonomy of the Psychomotor Domain. New York, David McKay Co.

Hecker, K. and C. Violato (2009). "Validity, reliability, and defensibility of assessments in veterinary education." J Vet Med Educ **36**(3): 271-275.

Highman, L., S. Marginson and V. Papatsiba (2023). "Higher education and research: multiple negative effects and no new opportunities after Brexit." Contemporary Social Science **18**(2): 216-234.

Holland, J. and N. Stevens (2020). Guidelines for the development of multiple choice items & assessments. Dublin, Ireland, RCSI University of Medicine and Health Sciences.

Hopkins, K. (1998). Educational and Psychological Measurement and Evaluation. Needham Heights, MA 02194;, Allyn & Bacon, A Viacom Company,.

HSE and NDTP (2023). NDTP Medical Workforce Planning for the Specialties of Pathology An Expert Stakeholder Informed Review. HSE.

Hulin, C. L., F. Drasgow and J. Komocar (1982). "Applications of item response theory to analysis of attitude scale translations." Journal of Applied Psychology **67**(6): 818.

Humphreys, H., E. Nagy, G. Kahlmeter and G. J. Ruijs (2010). "The need for European professional standards and the challenges facing clinical microbiology." Eur J Clin Microbiol Infect Dis **29**(6): 617-621.

Hurtz, G. and M. Auerbach (2003). "A Meta-Analysis of the Effects of Modifications to the Angoff Method on Cutoff Scores and Judgment Consensus." Educational and Psychological Measurement - EDUC PSYCHOL MEAS **63**: 584-601.

Hurtz, G. M. and N. R. Hertz (1999). "How Many Raters Should be Used for Establishing Cutoff Scores with the Angoff Method? a Generalizability Theory Study." Educational and Psychological Measurement **59**(6): 885-897.

Huynh, H. (2000). On item mappings and statistical rules for selecting binary

items for criterion-referenced interpretation and Bookmark standard settings. National Council on Measurement in

Education. New Orleans, LA.

Huynh, H. (2006). "A Clarification on the Response Probability Criterion RP67 for Standard Settings Based on Bookmark and Item Mapping." Educational Measurement: Issues and Practice **25**: 19-20.

Ibrahim, H., A. M. Juve, A. Amin, K. Railey and K. M. Andolsek (2023). "Expanding the Study of Bias in Medical Education Assessment." J Grad Med Educ **15**(6): 623-626.

Immekus, J., K. Snyder and P. Ralston (2019). "Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research." Frontiers in Education **4**.

Impara, J. C. and B. S. Plake (1998). "Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard Setting Method." Journal of Educational Measurement **35**(1): 69-81.

In, J. (2017). "Introduction of a pilot study." Korean J Anesthesiol **70**(6): 601-605.

Jaap, A., A. Dewar, C. Duncan, K. Fairhurst, D. Hope and D. Kluth (2021). "Effect of remote online exam delivery on student experience and performance in applied knowledge tests." BMC Medical Education **21**(1): 86.

Jaeger, R. M. (1982). "An Iterative Structured Judgment Process for Establishing Standards on Competency Tests: Theory and Application." Educational Evaluation and Policy Analysis **4**(4): 461-475.

Jaeger, R. M. (1991). "Selection of Judges for Standard-Setting." Educational Measurement: Issues and Practice **10**(2): 3-14.

Jennings, D. (2012). The Design of Multiple Choice Questions for Assessment (ABR). UCD Teaching and Learning. Ireland, UCD.

Jujo, S., J. J. Lee-Jayaram, B. I. Sakka, A. Nakahira, A. Kataoka, M. Izumo, K. Kusunose, N. Athinartrattanapong, S. Oikawa and B. W. Berg (2021). "Pre-clinical medical student cardiac point-of-care ultrasound curriculum based on the American Society of Echocardiography recommendations: a pilot and feasibility study." Pilot and Feasibility Studies **7**(1): 175.

Kane, M. (1994). "Validating the Performance Standards Associated With Passing Scores." Review of Educational Research **64**(3): 425-461.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. US, Lawrence Erlbaum. **5:** 129-145.

Karabekiroglu, K., B. Doğangün, S. Hergüner, T. von Salis and A. Rothenberger (2006). "Child and adolescent psychiatry training in Europe: differences and challenges in harmonization." Eur Child Adolesc Psychiatry **15**(8): 467-475.

Karantonis, A. and S. Sireci (2006). "The Bookmark Standard-Setting Method: A Literature Review." Educational Measurement: Issues and Practice **25**: 4-12.

Kass, M. E., J. M. Crawford, B. Bennett, T. M. Cox, M. M. Grimes, V. LiVolsi, C. D. Fletcher and D. S. Wilkinson (2007). "Adequacy of pathology resident training for employment: a survey report from the Future of Pathology Task Group." Arch Pathol Lab Med **131**(4): 545-555.

Kaufman, R. and J. M. Keller (1994). "Levels of evaluation: Beyond Kirkpatrick." Human Resource Development Quarterly **5**(4): 371-380.

Kelley, T. L. (1927). Interpretation of educational measurements. New York, Macmillan.

Kennedy, D., Á. Hyland and N. Ryan (2007). "Writing and Using Learning Outcomes: A Practical Guide."

Kerr, J. F. (1968). Changing the curriculum. London, University of London Press.

Kinley, K. (2023). Exploring alternative invigilated assessments in higher education: A work in progress.

Kirkpatrick, D. L. (1959). "Techniques for Evaluation Training Programs." journal of the American Society of Training Directors **13**: 21-26.

Knight, P. T. (2001). "Complexity and Curriculum: A process approach to curriculum-making." Teaching in Higher Education **6**(3): 369-381.

Kolen, M. and R. Brennan (2014). Test equating, scaling, and linking. Methods and practices. 3rd revised ed.

Kortese, L. (2016). "Exploring professional recognition in the EU: a legal perspective." Journal of international Mobility **4**(1): 43-58.

Krathwohl, D. R., B. S. Bloom and B. B. Masia (1964). Taxonomy of educational objectives.

Handbook II: Affective domain. New York, Longman.

Krstić, L., V. Aleksić and M. Krstić (2022). "Artificial intelligence in education: A review."

Kuder, G. F. and M. W. Richardson (1937). "The theory of the estimation of test reliability." Psychometrika **2**(3): 151-160.

Last, K., N. R. Power, S. Dellière, P. Velikov, A. Šterbenc, I. A. Antunovic, M. J. Lopes, V. Schweitzer and A. Barac (2021). "Future developments in training." Clin Microbiol Infect **27**(11): 1595-1600.

Laudonia, I., R. Mamlok-Naaman, S. Abels and I. Eilks (2018). "Action research in science education – an analytical review of the literature." Educational Action Research **26**: 480 - 495.

Lennon, M. C. (2018). Learning Outcomes Policies for Transparency: Impacts and Promising Practices in European Higher Education Regulation. European Higher Education Area: The Impact of Past and Future Policies. A. Curaj, L. Deca and R. Pricopie. Cham, Springer International Publishing**:** 527-546.

Lewis, D., H. Mitzel, R. Mercado and M. Schulz (2012). The Bookmark Standard Setting Procedure**:** 225.

Lewis, D. M., H. C. Mitzel and D. R. Green (1996). Standard setting: A Book-mark approach. In D. R. Green (Chair),IRT-based standard setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment. Phoenix, AZ.

Linacre, J. M. (2005). "Rasch dichotomous model vs. one-parameter logistic model." Rasch Measurement Transactions **19**(3): 1032.

Livingston, S. and M. Zieky (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests, Princeton, NJ: Educational Testing Service.

Lord, F. (1952). A theory of test scores (Psychometric Monograph No. 7). Iowa City, IA, Psychometric Society.

Low, D., S. W. Pollack, Z. C. Liao, R. Maestas, L. E. Kirven, A. M. Eacker and L. S. Morales (2019). "Racial/Ethnic Disparities in Clinical Grading in Medical School." Teach Learn Med **31**(5): 487-496.

Lowe, N. K. (2019). "What Is a Pilot Study?" J Obstet Gynecol Neonatal Nurs **48**(2): 117-118.

Lynch, D. and R. Smith (2011). The Theory and Practice of Curriculum and Programming**:** 30 to 42.

Mager, R. (1975). Preparing instructional objectives. Belmont, Pitman Learning.

Malhotra, S. D., K. N. Shah and V. J. Patel (2013). "Objective structured practical examination as a tool for the formative assessment of practical skills of undergraduate students in pharmacology." J Educ Health Promot **2**: 53.

Maraolo, A. E., D. S. Y. Ong, J. Cortez, K. Dedić, D. Dušek, A. Martin-Quiros, P. J. Maver, C. Skevaki, E. Yusuf, M. Poljak, M. Sanguinetti, E. Tacconelli, M. The Trainee Association of the European Society of Clinical and D. Infectious (2017). "Personal life and working conditions of trainees and young specialists in clinical microbiology and infectious diseases in Europe: a questionnaire survey." European Journal of Clinical Microbiology & Infectious Diseases **36**(7): 1287-1295.

Marceau, M., F. Gallagher, M. Young and C. St-Onge (2018). "Validity as a social imperative for assessment in health professions education: a concept analysis." Med Educ **52**(6): 641-653.

Mattick, K., R. Barnes and P. Dieppe (2012). "Medical education: A particularly complex intervention to research." Advances in health sciences education : theory and practice **18**.

Maurer, T. J. and R. A. Alexander (1992). "Methods of improving employment test critical scores derived by judging test content: A review and critique." Personnel Psychology **45**(4): 727-762.

McKernan, J. (1993). "Some Limitations of Outcome-Based Education." Journal of curriculum and supervision **8**.

McNamara, T. (2010). "The use of language tests in the service of policy: issues of validity." Revue française de linguistique appliquée **XV**(1): 7-23.

McNeir, G. (1993). Outcomes-based Education: Tool for Restructuring, Oregon School Study Council.

Mehtar, S. (1995). "Review of a consultant microbiologist's work practice--an audit." J Clin Pathol **48**(12): 1082-1086.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. Test validity. Hillsdale, NJ, US, Lawrence Erlbaum Associates, Inc**:** 33-48.

Messick, S. (1989). Educational measurement, 3rd ed, American Council on Education.

Messick, S. (1995). "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning." American Psychologist **50**(9): 741-749.

Miller, G. E. (1990). "The assessment of clinical skills/competence/performance." Acad Med **65**(9 Suppl): S63-67.

Mills, C. N. (1983). "A comparison of three methods of establishing cut-off scores on criterion-referenced tests." Journal of Educational Measurement **20**(3): 283-292.

Milone, A. S., A. M. Cortese, R. L. Balestrieri and A. L. Pittenger (2017). "The impact of proctored online exams on the educational experience." Curr Pharm Teach Learn **9**(1): 108-114.

Mitzel, H. C., D. M. Lewis, R. J. Patz and D. R. Green (2001). The bookmark procedure: Psychological perspectives. Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ, US, Lawrence Erlbaum Associates Publishers**:** 249-281.

Moore, C. G., R. E. Carter, P. J. Nietert and P. W. Stewart (2011). "Recommendations for planning pilot studies in clinical and translational research." Clin Transl Sci **4**(5): 332-337.

Mulder, L., A. Wouters, E. U. Akwiwu, A. S. Koster, J. H. Ravesloot, S. M. Peerdeman, M. Salih, G. Croiset and R. A. Kusurkar (2023). "Diversity in the pathway from medical student to specialist in the Netherlands: a retrospective cohort study." Lancet Reg Health Eur **35**: 100749.

Munna, A. and M. A. Kalam (2021). "Application of Theories, Principles and Models of Curriculum Design: A Literature Review." **3**: 147-153.

Musselwhite Thompson, D. and B. Wesolowski (2018). "Standard Error of Measurement."

Näsström, G. and P. Nyström (2008). "A comparison of two different methods for setting performance standards for a test with constructed-response items." Practical Assessment, Research and Evaluation **13**.

Neary, M. (2003). Curriculum concepts and research. In Curriculum studies in postcompulsory and adult education: A teacher's and student teacher's study guide. Cheltenham, Nelson Thornes Ltd.

Nedelsky, L. (1954). "Absolute Grading Standards for Objective Tests." Educational and Psychological Measurement **14**(1): 3-19.

Newble, D. (1998). Assessment. In Jolly B, Medical Education in the millenium. Oxford, Oxford University Press.

Norcini, J., B. Anderson, V. Bollela, V. Burch, M. J. Costa, R. Duvivier, R. Galbraith, R. Hays, A. Kent, V. Perrott and T. Roberts (2011). "Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference." Med Teach **33**(3): 206-214.

Norcini, J., M. B. Anderson, V. Bollela, V. Burch, M. J. Costa, R. Duvivier, R. Hays, M. F. Palacios Mackay, T. Roberts and D. Swanson (2018). "2018 Consensus framework for good assessment." Med Teach **40**(11): 1102-1109.

Norcini, J. and V. Burch (2007). "Workplace-based assessment as an educational tool: AMEE Guide No. 31." Med Teach **29**(9): 855-871.

Norcini, J. J. (2003). "Setting standards on educational tests." Med Educ **37**(5): 464-469.

Norcini, J. J. and J. A. Shea (1997). "The Credibility and Comparability of Standards." Applied Measurement in Education **10**(1): 39-59.

Norcini, J. J. J. (1999). "Standards and reliability in evaluation: when rules of thumb don't apply." Academic Medicine **74**(10): 1088-1090.

Notar, C. E., D. C. Zuelke, J. D. Wilson and B. D. Yunker (2004). "The Table of Specifications: Insuring Accountability in Teacher Made Tests." Journal of Instructional Psychology **31**: 115-129.

Nunnally, J. C. and I. H. Bernstein (1994). Psychometric Theory, McGraw-Hill.

O' Brien, J. (2008). "Writing Learning Outcomes A guide for academics Version 2." Retrieved 2024, 2024.

O'Neil, J. (1994). "Aiming for New Outcomes: The Promise and the Reality." Educational Leadership **51**(6): 6-10.

Olmos-Vega, F. M., R. E. Stalmeijer, L. Varpio and R. Kahlke (2023). "A practical guide to reflexivity in qualitative research: AMEE Guide No. 149." Medical Teacher **45**(3): 241-251.

Ong, D. S. Y., T. C. Zapf, M. Cevik, Z. R. Palacios-Baena, A. Barać, C. Cimen, A. E. Maraolo, C. Rönnberg, E. Cambau and M. Poljak (2019). "Current mentorship practices in the training of the next generation of clinical microbiology and infectious disease specialists: an international cross-sectional survey." Eur J Clin Microbiol Infect Dis **38**(4): 659-665.

Ornstein, A. and F. Hunkins (2009). Curriculum: Foundations, Principle and Theory. Boston, Allyn & Bacon.

Palacios-Baena, Z. R., T. C. Zapf, D. S. Y. Ong, A. E. Maraolo, C. Rönnberg, C. Çimen, C. Pulcini, J. Rodríguez-Baño, M. Sanguinetti, M. On behalf of the Trainee Association of the European Society of Clinical and D. Infectious (2018). "How are trainees in clinical microbiology and infectious diseases supervised in Europe? An international cross-sectional questionnaire survey by the Trainee Association of ESCMID." European Journal of Clinical Microbiology & Infectious Diseases **37**(12): 2381-2387.

Pamphlett, R. (2005). "It takes only 100 true–false items to test medical students: true or false?" Medical Teacher **27**(5): 468-470.

Pant, H., A. Rupp, S. Tiffin-Richards and O. Köller (2009). "Validity issues in standard-setting studies." Studies In Educational Evaluation **35**: 95-101.

Papapanou, M., E. Routsi, K. Tsamakis, L. Fotis, G. Marinos, I. Lidoriki, M. Karamanou, T. G. Papaioannou, D. Tsiptsios, N. Smyrnis, E. Rizos and D. Schizas (2022). "Medical education challenges and innovations during COVID-19 pandemic." Postgrad Med J **98**(1159): 321-327.

Pavlov, I. P. (1927). Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex, Oxford University Press, London, 142.

Pearson, N., P. Naylor, M. Ashe, M. Fernandez, S. L. Yoong and L. Wolfenden (2020). "Guidance for conducting feasibility and pilot studies for implementation trials." Pilot and Feasibility Studies **6**.

Peeling, R. W., D. L. Heymann, Y. Y. Teo and P. J. Garcia (2022). "Diagnostics for COVID-19: moving from pandemic response to control." Lancet **399**(10326): 757-768.

Perie, M. (2008). "A Guide to Understanding and Developing Performance-Level Descriptors." Educational Measurement: Issues and Practice **27**: 15-29.

Peytchev, A., F. G. Conrad, M. P. Couper and R. Tourangeau (2010). "Increasing Respondents' Use of Definitions in Web Surveys." J Off Stat **26**(4): 633-650.

Pfadenhauer, L. M., A. Gerhardus, K. Mozygemba, K. B. Lysdahl, A. Booth, B. Hofmann, P. Wahlster, S. Polus, J. Burns, L. Brereton and E. Rehfuess (2017). "Making sense of complexity in context and implementation: the Context and Implementation of Complex Interventions (CICI) framework." Implementation Science **12**(1): 21.

Phillips, J. (2012). Return on investment in training and performance improvement programs, second edition.

Piaget, J. (1952). Jean Piaget. A History of Psychology in Autobiography, Vol IV. Worcester, MA, US, Clark University Press**:** 237-256.

Pierse, T., R. Morris, O. T. L, B. Kinirons and E. Staddon (2023). "The retention of training doctors in the Irish health system." Ir J Med Sci **192**(6): 2573-2580.

Plake, B. S. (2005). Setting Performance Standards: Issues, Methods.

Polit, D. F. and C. T. Beck (2017). Nursing Research: Generating and Assessing Evidence for Nursing Practice, Wolters Kluwer Health.

Ponto, J. (2015). "Understanding and Evaluating Survey Research." J Adv Pract Oncol **6**(2): 168-171.

Quinn, F. M. (1997). The Principles and Practice of Nurse Education, Stanley Thornes.

Quinn, F. M. (2000). Principles and practice of nurse education. Cheltenham., Stanley Thornes (Publishers) Ltd.

Raoult, D., P. E. Fournier and M. Drancourt (2004). "What does the future hold for clinical microbiology?" Nature Reviews Microbiology **2**(2): 151-159.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, Danmarks Paedagogiske Institut.

Raymond, M. R. and J. P. Grande (2019). "A practical guide to test blueprinting." Medical Teacher **41**(8): 854-861.

RCPath(UK). (2020). "Autumn 2020,Online Examinations." 2024.

RCPath(UK). (2021). "Curriculum for speciality training in mEdical microbiology." Retrieved March 2024, 2024.

Reckase, M. (2006). "A Conceptual Framework for a Psychometric Theory for Standard Setting with Examples of Its Use for Evaluating the Functioning of Two Standard Setting Methods." Educational Measurement: Issues and Practice **25**: 4-18.

Reckase, M. D., M. J. Feuer and E. H. Haertel (2001). "The Controversy over the National Assessment Governing Board Standards." Brookings Papers on Education Policy(4): 231-265.

Reedy, A., D. Pfitzner, L. Rook and L. Ellis (2021). "Responding to the COVID-19 emergency: student and academic staff perceptions of academic integrity in the transition to online exams at three Australian universities." International Journal for Educational Integrity **17**(1): 9.

Reisenwitz, T. H. (2020). "Examining the Necessity of Proctoring Online Exams." Journal of Higher Education Theory and Practice **20**(1).

Rezigalla, A. A. (2015). "Angoff's Method: The Impact of Raters' Selection." Saudi Journal of Medicine & Medical Sciences **3**(3): 220-225.

Ricketts, C. (2009). "A plea for the proper use of criterion-referenced tests in medical assessment." Medical education **43**(12): 1141-1146.

Riordan, T., K. Cartwright, M. Logan, R. Cunningham, S. Patrick and T. Coleman (2002). "How do microbiology consultants undertake their jobs? A survey of consultant time and tasks in South West England." J Clin Pathol **55**(10): 735-740.

Roberts, C., S. Sarangi, L. e. Southgate, R. Wakeford and V. Wass (2000). "Oral examinations-equal opportunities, ethnicity, and fairness in the MRCGP." BMJ **320**: 370-375.

Royal, K. D. and M.-W. Hedgpeth (2015). "Balancing test length with sufficiently reliable scores." Education in Medicine Journal **7**(1).

Sabri, S. (2013). "Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities." International Journal of Education and Research **1**: 1-14.

Sackett, D. L. (1979). "Bias in analytic research." J Chronic Dis **32**(1-2): 51-63.

Sarrayrih, M. A. and M. Ilyas (2013). "Challenges of online exam, performances and problems for online university exam." International Journal of Computer Science Issues (IJCSI) **10**(1): 439-443.

Saylor, J. G., W. M. Alexander and A. J. Lewis (1981). Curriculum planning for better teaching and learning. New York, Holt, Rinehart & Winston.

Schön, D. A. (1987). Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions. San Francisco, CA, Jossey-Bass.

Schüttpelz-Brauns, K., E. Narciss, C. Schneyinck, K. Böhme, P. Brüstle, U. Mau-Holzmann, M. Lammerding-Koeppel and U. Obertacke (2016). "Twelve tips for successfully implementing logbooks in clinical training." Med Teach **38**(6): 564-569.

Schuwirth, L. W. and C. P. van der Vleuten (2004). "Different written assessment methods: what can be said about their strengths and weaknesses?" Med Educ **38**(9): 974-979.

Schuwirth, L. W. T. and C. P. M. van der Vleuten (2020). "A history of assessment in medical education." Adv Health Sci Educ Theory Pract **25**(5): 1045-1056.

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

Scott, G. M. (2000). "Clinical microbiology--the UK model." Clin Microbiol Infect **6**(8): 402-404.

Shanyinde, M., R. Pickering and M. Weatherall (2011). "Questions asked and answered in pilot and feasibility randomized controlled trials." BMC medical research methodology **11**: 117.

Shavelson, R. and N. Webb (1991). Generalizability Theory: A Primer.

Simmons, W. and L. Resnick (1993). "Assessment as the Catalyst of School Reform." Educational Leadership **50**(5): 11-15.

Sireci, S. (2009). "No More Excuses: New Research on Assessing Students with Disabilities." Journal of Applied Testing Technology **10**.

Skaggs, G. E. and A. A. Tessema (2001). Item Disordinality with the Bookmark Standard Setting Procedure.

Skinner, B. F. (1948). "'Superstition'in the pigeon." Journal of experimental psychology **38**(2): 168.

Skivington, K., L. Matthews, S. A. Simpson, P. Craig, J. Baird, J. M. Blazeby, K. A. Boyd, N. Craig, D. P. French, E. McIntosh, M. Petticrew, J. Rycroft-Malone, M. White and L. Moore (2021). "A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance." BMJ **374**: n2061.

Snadden, D. and M. L. Thomas (1998). "Portfolio learning in general practice vocational training--does it work?" Med Educ **32**(4): 401-406.

Socha, K. and J.-C. Dumont (2021). International migration and movement of doctors to and within OECD countries -2000 to 2018: Developments in countries of destination and impact on countries of origin.

Spady, W. G. (1988). "Organizing for Results: The Basis of Authentic Restructuring and Reform." Educational Leadership **46**: 4-8.

Spady, W. G. (1994). Outcome-Based Education: Critical Issues and Answers. Arlington, American Association of School Administrators.

Stenhouse, L. (1975). An introduction to Curriculum Research and Development. London, Heineman.

Swanson, D. B. and S. M. Case (1997). "Assessment in Basic Science Instruction: Directions for Practice and Research." Advances in Health Sciences Education **2**(1): 71-84.

Swanson, D. B. and C. P. van der Vleuten (2013). "Assessment of clinical skills with standardized patients: state of the art revisited." Teach Learn Med **25 Suppl 1**: S17-25.

Tavakol, M. and R. Dennick (2011). "Making sense of Cronbach's alpha." Int J Med Educ **2**: 53-55.

Tavakol, M. and R. Dennick (2013). "Psychometric evaluation of a knowledge based examination using Rasch analysis: an illustrative guide: AMEE guide no. 72." Med Teach **35**(1): e838-848.

Thabane, L., J. Ma, R. Chu, J. Cheng, A. Ismaila, L. P. Rios, R. Robson, M. Thabane, L. Giangregorio and C. H. Goldsmith (2010). "A tutorial on pilot studies: the what, why and how." BMC Medical Research Methodology **10**(1): 1.

Thomson, R. and G. Doern (2011). "What Will the Role of the Clinical Microbiology Laboratory Director Be in 2015?" J Clin Microbiol. **49**: S68-71.

Tiffin-Richards, S., H. Pant and O. Köller (2013). "Setting Standards for English Foreign Language Assessment: Methodology, Validation, and a Degree of Arbitrariness." Educational Measurement: Issues and Practice **32**.

Tighe, J., I. C. McManus, N. G. Dewhurst, L. Chis and J. Mucklow (2010). "The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations." BMC Med Educ **10**: 40.

Trigwell, K. (2001). "Judging university teaching." The International Journal for Academic Development **6**: 65-73.

Tummons, J. (2012). Curriculum studies in the lifelong learning sector. Exeter, Learning Matters.

Turnbull, J. M. (1989). "What is ... normative versus criterion-referenced assessment." Med Teach **11**(2): 145-150.

Tyler, R. W. (1949). Basic Principles of Curriculum and Instruction, University of Chicago Press.

UEMS (1993). Charter on Training of Medical Specialists. https://ebicm.esicm.org/assets/Upload/200536EU/Charter-for-training-1993.pdf.

UEMS (2013). Training Requirements for the

Specialty of Medical Microbiology. Medical Microbiology. https://uems-smm.eu/application/files/3016/4210/5793/European_Training_Requirements_in_Medical_Microbiology_-_September_2013.pdf.

UEMS (2017). CORE TRAINING PROGRAMME FOR MEDICAL MICROBIOLOGY. section Medical Microbiology. https://uems-smm.eu/application/files/3416/4210/5793/4.2.7-UEMS-MM-curriculum-submitted-European-council-2017.09.21.pdf, UEMs Section Medical Microbiology.

Umobong, M. E. (2017). "The One-Parameter Logistic Model (1PLM) And Its Application In Test Development." Advances in Social Sciences Research Journal **4**.

van der Aa, J. E., F. Scheele, A. J. Goverde and P. W. Teunissen (2019). "A qualitative study on harmonization of postgraduate medical education in Europe: negotiating flexibility is key." Perspect Med Educ **8**(4): 216-222.

van der Vleuten, C. and B. Verhoeven (2013). "In-training assessment developments in postgraduate education in Europe." ANZ J Surg **83**(6): 454-459.

Van der Vleuten, C. P. and D. B. Swanson (1990). "Assessment of clinical skills with standardized patients: State of the art." Teaching and Learning in Medicine **2**(2): 58-76.

Van Der Vleuten, C. P. M. (1996). "The assessment of professional competence: Developments, research and practical implications." Advances in Health Sciences Education **1**(1): 41-67.

van Teijlingen, E. and V. Hundley (2002). "The Importance of Pilot Studies." Nursing standard (Royal College of Nursing (Great Britain) : 1987) **16**: 33-36.

Wang, N. (2003). "Use of the Rasch IRT Model in Standard Setting: An Item-mapping Method." Journal of Educational Measurement **40**: 231-253.

Ward, R. C., T. J. Muckle, M. J. Kremer and M. A. Krogh (2019). "Computer-Based Case Simulations for Assessment in Health Care: A Literature Review of Validity Evidence." Evaluation & the Health Professions **42**(1): 82-102.

Wass, V., C. Van der Vleuten, J. Shatzer and R. Jones (2001). "Assessment of clinical competence." Lancet **357**(9260): 945-949.

Wass, V., R. Wakeford, R. Neighbour and C. Van der Vleuten (2003). "Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component." Med Educ **37**(2): 126-131.

Watson, J. B. and R. Rayner (1920). "Conditioned emotional reactions." Journal of Experimental Psychology **3**(1): 1-14.

Wayne, D. B., M. J. Fudala, J. Butter, V. J. Siddall, J. Feinglass, L. D. Wade and W. C. McGaghie (2005). "Comparison of two standard-setting methods for advanced cardiac life support training." Acad Med **80**(10 Suppl): S63-66.

Westerling, R. (2009). "The harmonization of the medical speciality in Public Health in the EU countries—a challenge for the profession." European Journal of Public Health **19**(3): 230-232.

Wiliam, D. (1996). "Meanings and Consequences in Standard Setting." Assessment in Education: Principles, Policy & Practice **3**(3): 287-308.

Wiliam, D. (2008). Quality in assessment**:** 123-137.

Winzor, G. and M. Patel (2015). "Combined infection training - should we be concerned about its impact on infection prevention and control training of microbiologists in the UK?" J Hosp Infect **91**(4): 302-305.

Wright BD and L. JM. (1994). "Rasch Measurement Transactions."

Wu, M.-J., K. Zhao and F. Fils-Aime (2022). "Response rates of online surveys in published research: A meta-analysis." Computers in Human Behavior Reports **7**: 100206.

Yigit, C. and B. Bagceci (2017). "Teachers' Opinions Regarding the Usage of Action Research in Professional Development." Journal of Education and Training Studies **5**: 243.

Yim, M. K. and S. Shin (2020). "Using the Angoff method to set a standard on mock exams for the Korean Nursing Licensing Examination." Journal of Educational Evaluation for Health Professions **17**.

Yousefi Afrashteh, M. (2021). "Comparison of the validity of bookmark and Angoff standard setting methods in medical performance tests." BMC Med Educ **21**(1): 1.

Yun, G. W. and C. Trumbo (2000). "Comparative Response to a Survey Executed by Post, E-Mail, and Web Form." J. Computer-Mediated Communication **6**.

Yusuf, E., D. S. Ong, A. Martin-Quiros, C. Skevaki, J. Cortez, K. Dedić, A. E. Maraolo, D. Dušek, P. J. Maver, M. Sanguinetti and E. Tacconelli (2017). "A large survey among European trainees in clinical microbiology and infectious disease on training systems and training adequacy: identifying the gaps and suggesting improvements." Eur J Clin Microbiol Infect Dis **36**(2): 233-242.

Zarei, M., H. Eftekhari Mamaghani, A. Abbasi and M.-S. Hosseini (2024). "Application of artificial intelligence in medical education: A review of benefits, challenges, and solutions." Medicina Clínica Práctica **7**(2): 100422.

Zieky, M. and S. Livingston (1977). Basic Skills Assessment. Manual for Setting Standards on the Basic Skills Assessment Tests. Princeton, NJ., Educational Testing Service,.

Zuber-Skerritt, O. and M. Fletcher (2007). "The quality of an action research thesis in the social sciences." Quality Assurance in Education **15**: 413-436.

Zumbo, B. (2016). "Standard-setting methodology: Establishing performance standards and setting cut scores to assist score interpretation." Applied Physiology Nutrition and Metabolism **vol. 41, (Number, 6; Suppl. 2), S74-S82.**